Journal of Machine Learning Research 24 (2023) 1-56 Submitted 11/20; Revised 12/21, 11/22; Published 4/23

Evaluating Instrument Validity using the Principle of Independent Mechanisms

Patrick F. Burauel PBURAUEL@CALTECH.EDU
California Institute of Technology
Pasadena, CA, USA

Editor: Joris Mooij

Abstract

The validity of instrumental variables to estimate causal effects is typically justified narratively and often remains controversial. Critical assumptions are difficult to evaluate since they involve unobserved variables. Building on Janzing and Schölkopf's (2018) method to quantify a degree of confounding in multivariate linear models, we develop a test that evaluates instrument validity without relying on Balke and Pearl's (1997) inequality constraints. Instead, our approach is based on the Principle of Independent Mechanisms, which states that causal models have a modular structure. Monte Carlo studies show a high accuracy of the procedure. We apply our method to two empirical studies: first, we can corroborate the narrative justification given by Card (1995) for the validity of college proximity as an instrument for educational attainment in his work on the financial returns to education. Second, we cannot reject the validity of past savings rates as an instrument for economic development to estimate its causal effect on democracy (Acemoglu et al., 2008).

Keywords: instrumental variables, Principle of Independent Mechanisms, causality, unobserved confounding, causal inference from observational data

1. Introduction

Scientific analysis often seeks to provide estimates for the causal effects of variables under study. Concerns about unobserved confounding, which can invalidate such estimates, are widespread in non-experimental studies in many disciplines such as economics and epidemiology. To estimate a causal effect in spite of unobserved confounding, a common solution is to resort to instrumental variable (IV) techniques.

A typical IV setting is depicted in Figure 1: A treatment variable T has a causal effect τ on an outcome Y. In addition, there is an unobserved confounder U, which influences both T and Y. A naive estimate of τ based on statistical correlation between observed T and Y would contain a mixture of the true causal effect and the confounding effect. An instrumental variable, or simply *instrument*, Z can help to disentangle the causal and confounded parts if it satisfies critical IV assumptions: 1) The instrument must be statistically related to the treatment variable T (relevance). 2) The relation of the instrument to the outcome Y must not be confounded, that is, there must not be an unobserved variable that influences both Z and Y (exchangeability assumption). 3) The instrument must not be a direct cause of the effect Y, rather it must have a causal effect on Y only indirectly via the

©2023 Patrick Burauel.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v24/20-1287.html.

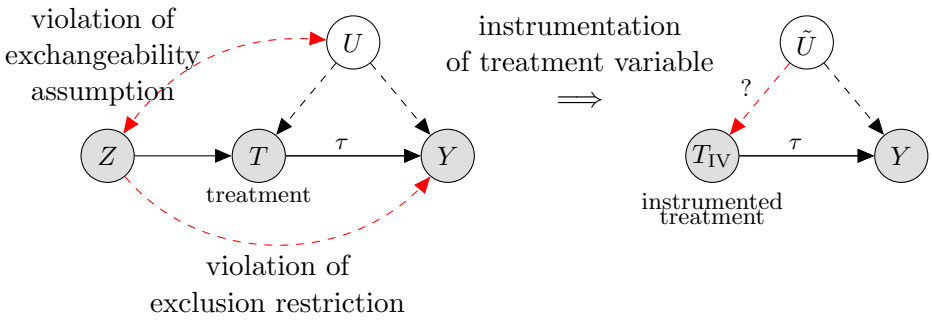

Figure 1: *Graphical representation of an illustrative Instrumental Variable model. Dark nodes represent observed variables, light nodes represent unobserved variables. Left panel: T represents the treatment variable of interest, which has a causal effect (τ) on Y. An instrumental variable (Z) can help to identify τ in spite of unobserved confounders (U) if it does not have a direct causal effect on Y (exclusion restriction) and is not related to unobserved confounder U (exchangeability assumption). The red arrow from Z to Y indicates how the exclusion restriction can be violated by Z's direct effect on Y. The double-edged arrow between U and Z indicates how the exchangeability assumption can be violated when there is an unobserved confounder influencing both Z and Y. Right panel: If either of the two arrows is present, the instrument is not valid and using it to instrument T with Z yields a instrumented treatment variable (T_{IV}) whose relation to Y is confounded. To emphasize that the unobserved variable in the right panel potentially confounds T_{IV} and Y (and not T and Y) we denote it with $\tilde{U}$ (and not U).*

treatment T (exclusion restriction). An IV that fulfills these assumptions is called "valid". See Section 3.1 for more details.

A valid IV can be used to extract experimental (or exogeneous) variation in T that is unrelated to the confounder U. In other words, it can be used to construct an instrumented treatment variable T_{IV} that is unconfounded with Y. It is then possible to use T_{IV} to get a consistent estimate of the sought causal effect τ.[1] However, it is difficult to know whether the IV assumptions are satisfied. In particular, the validity of exclusion restriction and exchangeability assumption are difficult to evaluate because they involve unobserved variables. The right panel of Figure 1 illustrates what the problem of IV validity boils down to: does instrumenting the treatment variable lead to an instrumented treatment variable that is unconfounded? The method proposed here is able to test whether T_{IV} is confounded with Y or not. Since T_{IV} is unconfounded if the instrument is valid, the method can indirectly evaluate critical IV assumptions.

In practice, scientists need to rely on narrative and often controversial justifications of those critical validity assumptions. Therefore, it is important to develop and make accessible statistical tests that can falsify IV validity. Consider an example from economics. Acemoglu et al. (2008) ask whether economic development causes democratic development—a relation that is likely confounded by a plethora of variables such as the level of general education. They estimate a causal effect of {economic development} on the degree of {democratic

1. In settings where the instrument and treatment are binary, the true causal effect cannot be point-identified but can be bounded (Labrecque and Swanson, 2018).

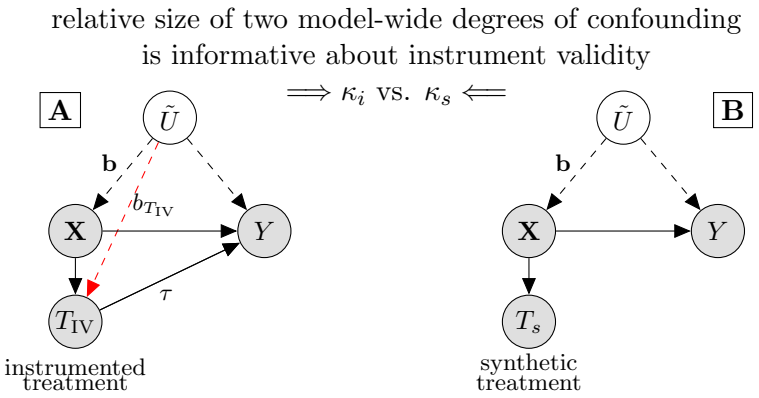

Figure 2: *Illustration of the proposed test for instrument validity. Panel A: Instrumental variable model with additional covariates* $\mathbf{X}$*. The instrumented treatment variable* T_{IV} *is confounded* ($b_{T_{IV}} \neq 0$) *if the instrument is invalid, cf. Figure 1. The degree of confounding of the IV model is denoted* κ_i*. Panel B: The instrumented treatment is replaced by a synthetic variable* T_s *that is unconfounded by construction* ($b_{T_{IV}} = 0$)*. The degree of confounding of this counterfactual model is denoted* κ_s*. The relative size of* κ_i *and* κ_s*, two quantities that can be estimated from observed data, is informative about instrument validity. This result holds in spite of* T_s *not having a causal effect on* Y*. See detailed discussion in Section 3.*

development} by using {past saving rates} as an instrument for {economic development}. The exclusion restriction demands that {past saving rates} do not have a direct effect on {democratic development}. In defense of that instrument, they argue that "it seems *plausible* to expect that changes in the savings rate over periods of five to ten years should have no direct effect on the culture of democracy" (p. 822, italics added). This example illustrates how the justification of critical assumptions in IV studies is typically not substantiated by sound statistical tests[2] and, thus, underscores the need to develop such tests.

To develop such a test for IV validity, we build on a method to estimate a degree of confounding in multivariate linear models proposed by Janzing and Schölkopf (2018a) (which is denoted as JS throughout). Their method rests on the Principle of Independent Mechanisms—succinctly, a model that represents causal relations has a modular structure. JS propose a way to estimate a degree of confounding, $\kappa \in [0, 1]$, for a linear model where one outcome variable, Y, is correlated with a high-dimensional set of potential causes, $\mathbf{X}$. κ measures the extent to which the observed correlation between high-dimensional $\mathbf{X}$ and Y is due to genuine causation, $\kappa = 0$ (no confounding, all observed statistical relation is due to causation), or due to a confounder, $\kappa = 1$ (full confounding, all observed statistical relation is due to confounding).

Since κ measures the extent to which the whole set of high-dimensional $\mathbf{X}$ is confounded and instrument validity is about a single potentially confounded variable, the off-the-shelf version of JS is not applicable to evaluate instrument validity. We solve that problem by

2. Acemoglu et al. (2008) augment their *plausability* argument by controlling for a number of additional covariates and checking whether the coefficient of interest changes. Yet, this is shown to be an uninformative procedure in observational studies (Oster, 2019). Further, the authors employ an overidentification test, which *assumes* validity of at least one instrument.

providing a way to estimate a counterfactual degree of confounding κ_s that would be obtained if the instrument were valid. It relies on generating a variable that is unconfounded by construction and yet "similar" to the possibly confounded instrumented treatment variable. The counterfactual degree of confounding κ_s can be compared to the actual κ_i that is obtained from the IV model under consideration. We show that the difference between κ_s and κ_i is informative about instrument validity. This result allows for testing IV validity. See Figure 2 for an illustration. Unlike other tests (described in Section 2), the proposed test benefits from the presence of high-dimensional control covariates.

Section 2 provides an overview of related research. In Section 3, the core of this paper, we discuss how to estimate the counterfactual degree of confounding and how to use it to test IV validity. While we take JS as given in the contruction of our test, Section 4 provides a more detailed discussion of the Principle of Independent Mechanisms, introduces the JS method and conveys graphical intuition for its functioning. Section 5 provides Monte Carlo simulation studies, which show high accuracy of the methodology. Section 6 contains two empirical applications. First, we apply the proposed method to a study by Card (1995), who proposes to use {college proximity} of a family's residence as an instrument for {educational attainment} to estimate {financial returns to education}. Though Card himself casts doubt on the validity of {college proximity} as an instrument, he argues that the instrument is likely valid in specific subsamples of the data. The proposed methodology corroborates Card's narrative justification of the validity of the instrument in specific subsamples. Second, we cannot reject the validity of {past saving rates} as an instrument for {economic development} in a study on the causes of {democratic development} by Acemoglu et al. (2008).

2. Previous Research

The Sargan (1958)-Hansen (1982) J-test for overidentifying restrictions arguably spawned the substantial literature on specification testing in instrumental variable (IV) models. The J-test can be used to test instrument validity when there are more instruments than possibly confounded treatment variables. Failure of rejecting the null hypothesis of the J-test is evidence that all proposed instruments are valid. Rejecting the null provides evidence that at least one of the proposed instruments is invalid. However, the test cannot determine which of the proposed instruments is invalid and, therefore, is not useful to choose a subset of valid instruments.

A more recent strand of the literature proposes nonparametric tests for unconfoundedness of explanatory variables, e.g. an instrumented treatment variable. In broad terms, what unites many of these papers is their reliance on testing whether the moment conditions implied by the instrumental variable model are fulfilled. By analyzing higher-order moments, these models can resort to overidentifying restrictions even when there is only one instrument per confounded variable. For example, Blundell and Horowitz (2007) propose a test for unconfoundedness in nonparametric regression analysis that does not rely on nonparametric IV estimation (which often suffers from slow convergence that, in turn, results in low power of such tests). Two related papers that both study nonparametric IV models are Breunig (2015) and Gagliardini and Scaillet (2017). The former uses series estimators to propose a test for instrument validity and the latter employ a Tikhonov-regularized es-

timator of the functional parameter to minimize the distance criterion corresponding to the moment conditions. Breunig (2018) extends these results to nonparametric quantile regression with nonseparable errors.

Although diverse methods to test IV validity in *overidentified* IV models are proposed, those for just-identified models prove more elusive. However, the causal structure in IV models with binary instrument and binary treatment ("binary IV models") implies testable constraints on the outcome distribution of four groups of individuals defined by two observed quantities (treatment status and instrument assignment). Those are described by Balke and Pearl (1997). Specifically, if the outcome distributions of individuals with $Z_i = 1,\ T_i = 0$ and $Z_i = 0,\ T_i = 0$, or those of individuals with $Z_i = 1,\ T_i = 1$ and $Z_i = 0,\ T_i = 1$ intersect, instrument validity is violated. These testable constraints are first leveraged by Kitagawa (2015), who proposes to test instrument validity by checking whether the aforementioned distributions intersect. Huber and Mellace (2015) provide a closely related extension to Kitagawa (2015): they propose a test that relies on mean potential outcomes rather than their distributions. Mourifié and Wan (2017) build on Kitagawa (2015) by representing his test in terms of conditional moment inequalities.

Kitagawa's work and the mentioned extensions are applicable in binary IV models since that is the context in which Balke and Pearl's testable constraints arise. This paper provides a test to detect invalid instruments that does not use Balke and Pearl's testable implications. Instead, it takes another angle at the problem of evaluating IV validity: it relies on the idea that invariance structures in observed data justify statements about the underlying causal structure of the system under study. This idea is formalized from an information-theoretic perspective as the Principle of Independent Mechanisms (PIM) (Janzing et al., 2012; Peters et al., 2017, and is discussed further in Section 4). Since it does not rely on Balke and Pearl's testable constraints, the method proposed here is not restricted to binary IV models. Unlike Kitagawa (2015) whose test is applicable in nonparametric models, we are restricted to linear models in this paper. This is because we use an interpretation of PIM for linear models. Note that Balke and Pearl (1997) prove that their testable implications are sharp in the sense that there is no additional restriction on the data distribution that could be used to test IV validity. It is shown by Heckman and Vytlacil (2005) that additional distributional restrictions result from the constant treatment effects model that we are considering. Therefore, our restriction is not sharp. The focus of this paper is to show that a restriction resulting from a formalization of the Principle of Independent Mechanisms, which is a type of restriction that has so far not been applied, enables testing IV validity.

In sum, the main contribution of this paper is to develop a novel testing approach for instrument validity that relies neither on moment restrictions nor on Balke and Pearl's testable implications. Our approach is based, intuitively, on the Principle of Independent Mechanisms and, technically, on the decomposition of the spectral measure of the covariates' covariance matrix induced by the (possibly biased) corresponding parameter vector. By assuming PIM, our approach makes a structural assumption that has hitherto not been used to evaluate instrument validity (and that is not necessary to *estimate* the sought causal effect). Using PIM in this context only becomes possible through its recent formalization due to Janzing and Schölkopf (2018a). The present work adds to the growing literature using PIM as a powerful concept to guide causal identification (see e.g. Peters et al., 2016; Besserve et al., 2018a,b; Gresele et al., 2021).

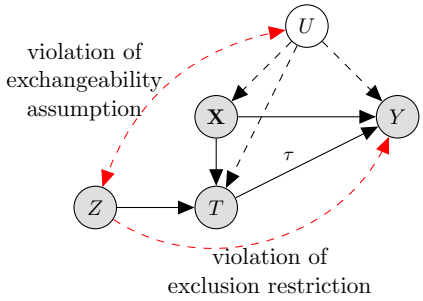

Figure 3: *Graphical representation of the IV model under study. The difference between the IV model under study and the illustrative IV model depicted in Figure 1 is the presence of additional control variables* **X**.

3. Test for Instrument Validity

In this section, we formally define the model under consideration, discuss assumptions for instrument validity and describe the test procedure to evaluate instrument validity. We take the method to estimate a degree of confounding in multivariate linear models proposed by Janzing and Schölkopf (2018a) (JS) as given and mention details only insofar as they are relevant for developing the test. Since it is an integral part of the test proposed here, we describe JS informally in Section 4 and formally in Appendix H.

3.1 Assumptions for instrument validity

Figure 3 shows a graphical representation of the conditional instrumental variable model under consideration. The difference between this model and the illustrative IV model in Figure 1 (discussed in the Section 1) is the presence of additional control variables **X**, which have causal effects on the treatment variable T and the outcome Y. Before introducing the parametric version of this model in the next subsection, we discuss the assumptions that an instrument Z has to fulfill to enable the consistent estimation of τ in spite of unobserved U (see Labrecque and Swanson, 2018; Didelez et al., 2010).

Assumption 1. *relevance assumption*

$$Z \not\perp\!\!\!\perp T \tag{1}$$

The relevance assumption states that the instrument must be statistically related to the treatment variable T. This assumption can easily be tested by, e.g., checking whether a regression of T on Z produces a coefficient estimate that is significantly different from zero. Since it is easy to test, this assumption will be taken for granted in the following.

Assumption 2. *exchangeability assumption*

$$Z \perp\!\!\!\perp U \text{ and } Z \perp\!\!\!\perp Y | T, U. \tag{2}$$

Assumption 2 implies an absence of an unobserved confounder between Z and U as well as between Z and Y. Note that the bi-directed dashed red arrow between Z and Y, which would violate $Z \perp\!\!\!\perp Y | T, U$, is not depicted in Figure 3.

Assumption 3. *exclusion restriction*

$$Z \perp\!\!\!\perp Y | do(T = t), \mathbf{X}. \tag{3}$$

The $do(V = v)$ operator denotes an intervention on variable V that sets it to v and deletes all incoming edges to V (Pearl, 2009). The exclusion restriction implies the absence of a direct causal effect of Z on Y.

With these assumptions we can provide a definition of IV validity:

Definition 1. *A variable Z is called a valid instrumental variable if and only if it fulfills Assumptions 1 to 3.*

Figure 3 illustrates violations of these crucial IV assumptions with the dashed red arrows. A bi-directed edge denotes an unobserved confounder between the respective nodes. A directed edge represents a direct causal relation.

3.2 Parametric latent IV model

We consider the following structural linear IV model with a constant treatment effect τ and an additively separable error term.

$$Y = \mathbf{X}\beta + \tau T + \beta_u U + \varepsilon_Y \text{ and} \tag{4}$$

$$T = \mathbf{X}\gamma + \gamma_z Z + \gamma_u U + \varepsilon_T \tag{5}$$

where $\mathbf{X}$ represents a set of d covariates, T is a treatment variable, and Z is a binary instrument. Y is the outcome variable of interest. U is an unobserved confounder. β and γ are d-dimensional vectors of coefficients. β_u, γ_z, and γ_u are scalar coefficients. τ is the scalar causal effect of interest: the Average Treatment Effect (ATE). ε_Y and ε_T are structural error terms, which are independent of each other. We refer to the assumed model structure in eqs. (4) and (5) as the Maintained Assumption. Unlike the model studied by Angrist et al. (1996), this model does not allow for heterogeneous treatment effects. We assume a constant treatment effect, see discussion in Section 3.7.

3.3 Reduced form model and test idea

A common estimator for the causal effect of interest in IV models is the two-stage least squares estimator (2SLS), as discussed by e.g. Wooldridge (2002). Its implementation comprises two steps. First, the treatment variable is regressed on the instrument and additional control variables. This regression is used to calculate predictions for T, which we denote with T_{IV}. Second, the outcome variable Y is regressed on T_{IV} from the first stage and the additional control variables, that is Y is regressed on $\{\mathbf{X}, T_{\text{IV}}\}$. The coefficient of T_{IV} is a consistent estimate of the causal effect τ if the instrument is valid. To develop the test, we reformulate the model in eqs. (4)-(5) to its reduced form after the first stage is implemented. After replacing the observed treatment variable with its instrumented version, the confounding variable we are concerned with is no longer U (which confounds $\{\mathbf{X}, T\}$ and Y) but a different one, namely that variable which confounds $\{\mathbf{X}, T_{\text{IV}}\}$ and Y,

which we call $\tilde{U}$. To parameterize the degree of condounding of each control variable and T_{IV}, we express them as a sum of their unconfounded versions and the scalar confounder $\tilde{U}$:

$$Y = \{\mathbf{X}, T_{\text{IV}}\}\mathbf{a} + c\tilde{U} + \varepsilon \text{ and} \tag{6}$$

$$\{\mathbf{X}, T_{\text{IV}}\} = \mathbf{E} + \tilde{U}\begin{pmatrix}\mathbf{b} & b_{T_{\text{IV}}}\end{pmatrix} \tag{7}$$

where $\{\mathbf{X}, T_{\text{IV}}\}$ denotes a matrix of d control variables $\mathbf{X}$ and the instrumented treatment variable T_{IV}. $\mathbf{a} = \begin{pmatrix}\beta \\ \tau\end{pmatrix}$ where β is the d-dimensional parameter vector associated with covariates $\mathbf{X}$ and τ is the true causal effect of interest. $\tilde{U}$ is an unobserved confounder, which influences Y when $c \neq 0$ and $\{\mathbf{X}, T_{\text{IV}}\}$ when $\begin{pmatrix}\mathbf{b} & b_{T_{\text{IV}}}\end{pmatrix} \neq \mathbf{0}$. The hypothetical unconfounded versions of $\mathbf{X}$ and T_{IV} are represented by $\mathbf{E} = \{\mathbf{X}^*, T_{\text{IV}}^*\}$. Confounding is introduced by adding $\tilde{U}\begin{pmatrix}\mathbf{b} & b_{T_{\text{IV}}}\end{pmatrix}$. Thus, each element of the vector $\begin{pmatrix}\mathbf{b} & b_{T_{\text{IV}}}\end{pmatrix} = \begin{pmatrix}b_1 & \dots & b_d & b_{T_{\text{IV}}}\end{pmatrix}$ parameterizes confounding of the corresponding dimension of $\{\mathbf{X}, T_{\text{IV}}\}$, e.g. $X_1 = E_1 + \tilde{U}b_1$.

If Z is a valid IV (it can be used to estimate τ consistently), then the T_{IV} resulting from 2SLS must be unconfounded. Specifically, the element $b_{T_{\text{IV}}}$ that parameterizes the confounding of the instrumented treatment variable T_{IV} will be zero (cf. $T_{\text{IV}} = T_{\text{IV}}^* + \tilde{U}b_{T_{\text{IV}}}$) if the instrument is valid. To foreshadow, the proposed procedure enables evaluating whether $b_{T_{\text{IV}}} = 0$. Since any violation of Assumption 2 or Assumption 3 leads to a confounded T_{IV} (that is, $b_{T_{\text{IV}}} \neq 0$), the test is not capable of disambiguating *which* of the two assumptions invalidates the instrument even though it is capable of detecting both violations.

So, how to evaluate whether $b_{T_{\text{IV}}} = 0$? First, we need to take a step a step back to understand the definition of the level of confounding that JS show how to estimate, which is defined as[3]

$$\kappa := \frac{\|\hat{\mathbf{a}} - \mathbf{a}\|^2}{\|\mathbf{a}\|^2 + \|\hat{\mathbf{a}} - \mathbf{a}\|^2} \tag{8}$$

$$= \frac{\left\| c\Sigma^{-1}_{\mathbf{X}T_{\text{IV}}} \begin{pmatrix}\mathbf{b} \\ b_{T_{\text{IV}}}\end{pmatrix} \right\|^2}{\left\| \begin{pmatrix}\beta \\ \tau\end{pmatrix} \right\|^2 + \left\| c\Sigma^{-1}_{\mathbf{X}T_{\text{IV}}} \begin{pmatrix}\mathbf{b} \\ b_{T_{\text{IV}}}\end{pmatrix} \right\|^2} \in [0, 1] \tag{9}$$

where $\hat{\mathbf{a}} = \Sigma^{-1}_{\mathbf{X}T_{\text{IV}}}\Sigma_{\mathbf{X}T_{\text{IV}}Y}$ denotes the parameter vector after projecting with least-squares in the population, that is, $\Sigma_{\mathbf{X}T_{\text{IV}}}$ is the covariance matrix of $\{\mathbf{X}, T_{\text{IV}}\}$ and $\Sigma_{\{\mathbf{X}T_{\text{IV}}\}Y}$ is the covariance vector of $\{\mathbf{X}, T_{\text{IV}}\}$ with Y. $c\Sigma^{-1}_{\mathbf{X}T_{\text{IV}}} \begin{pmatrix}\mathbf{b} \\ b_{T_{\text{IV}}}\end{pmatrix} = \hat{\mathbf{a}} - \mathbf{a}$, which is the deviation of the parameter vector $\hat{\mathbf{a}}$ from the structural parameter $\mathbf{a}$. To get an intuitive understanding what κ measures, consider the approximation $\|\hat{\mathbf{a}}\|^2 \approx \|\hat{\mathbf{a}} - \mathbf{a}\|^2 + \|\mathbf{a}\|^2 +$. It holds when $\hat{\mathbf{a}} - \mathbf{a}$ is orthogonal to $\mathbf{a}$, which is approximately true as the dimensionality of these vectors goes to infinity (see Janzing and Schölkopf, 2018a, see also eq. (73) in Appendix H). Using this approximation, one can see that

$$\kappa \approx \frac{\|\hat{\mathbf{a}} - \mathbf{a}\|^2}{\|\hat{\mathbf{a}}\|^2}. \tag{10}$$

3. Throughout, $\|a\|$ denotes the L_2 norm of the d-dimensional vector a.

Therefore, κ approximately equals the deviation of $\hat{\mathbf{a}}$ from $\mathbf{a}$ relative to $\hat{\mathbf{a}}$, in terms of the respective squared lengths of these vectors. Since the degree of confounding is defined in terms of squared lengths of vectors, it is impossible to trace confounding back to single elements of these vectors in the original JS method.

The aggregate nature of κ is an important limitation given that confounding of a single covariate is what is informative about instrument validity. Recall that κ lies between 0 and 1. Confounding is introduced by vector $\mathbf{b}_{\text{IV}} = \begin{pmatrix} \mathbf{b} \\ b_{T_{\text{IV}}} \end{pmatrix}$ and scalar c. The degree of confounding is zero ($\kappa = 0$) when $\mathbf{b}_{\text{IV}} c = \mathbf{0}$ and it is positive ($\kappa > 0$) when $\mathbf{b}_{\text{IV}} c \neq \mathbf{0}$. Thus, the estimable κ sheds light on the product of two unobserved quantities $\mathbf{b}_{\text{IV}}$ and c. This underlines that observing $\kappa > 0$ is not informative about *which dimension* of the covariates $\{\mathbf{X}, T_{\text{IV}}\}$ is confounded. In other words, $\kappa_i := \kappa(\{\mathbf{X}, T_{\text{IV}}\}, Y)$ gives an overall degree of confounding of the model where Y is regressed on $\{\mathbf{X}, T_{\text{IV}}\}$, without specifying which specific dimensions are confounded. To deduce whether the instrumental variable Z is valid, however, it is essential to know whether a specific covariate, namely T_{IV}, is confounded or not, that is whether $b_{T_{\text{IV}}} = 0$ or not. However, a single κ estimate is not informative about $b_{T_{\text{IV}}}$.

To address this problem, we propose a way to estimate confounding of a *single* variable that builds on JS. We do this by estimating a counterfactual degree of confounding, κ_s, that would be obtained if that single covariate T_{IV} were unconfounded. This is achieved by generating a synthetic treatment variable T_s that is unconfounded by construction. Replacing T_{IV} with T_s, results in the synthetic reduced form model:

$$Y = \{\mathbf{X}, T_s\} \begin{pmatrix} \beta \\ \tau_s \end{pmatrix} + c\tilde{U} + \varepsilon \text{ and} \tag{11}$$

$$\{\mathbf{X}, T_s\} = \mathbf{E} + \tilde{U} \begin{pmatrix} \mathbf{b} & b_{T_s} \end{pmatrix}. \tag{12}$$

Note that the element corresponding to T_s in the vector that multiplies the confounder is equal to zero because T_s is unconfounded by construction: $b_{T_s} = 0$. Also, since the synthetic T_s does not have a causal effect on Y, $\tau_s = 0$.

In sum, we replace T_{IV} by the synthetic (and unconfounded) T_s to estimate $\kappa_s := \kappa(\{\mathbf{X}, T_s\}, Y)$ (the degree of confounding that would be obtained with a valid IV). Then, we compare this counterfactual κ_s to the actual degree of confounding κ_i. We can show that their relative size is informative about instrument validity. In the following Section, we discuss the details of this idea and its implementation.

3.4 Detailed test procedure

The test procedure is succinctly described in Algorithm 1. In the following main text we provide a description that focuses on the intuition. We denote the degree of confounding as measured by the method laid out in Janzing and Schölkopf (2018a) (JS) in a multivariate linear model with $\mathbf{X}$ as independent variables and Y as dependent variable with $\kappa(\{\mathbf{X}\}; Y)$.

Under the Maintained Assumption, which is the model structure assumed in eqs. (4)-(5), we want to test the hypothesis

$$H_0 : Z \text{ is a valid instrument} \tag{13}$$

against the alternative

$$H_1 : Z \text{ is not a valid instrument.} \tag{14}$$

To indirectly test IV validity by analyzing the extent to which the instrumented treatment variable (T_{IV}) is confounded (invalid IV) or unconfounded (valid IV), we adapt the method to estimate a model-wide degree of confounding due to JS.

The degree of confounding of the model that contains the instrumented treatment variable T_{IV} (that is, the degree of confounding of the model depicted in Panel A of Figure 2) is denoted

$$\kappa_i := \kappa(\{\mathbf{X}, T_{\text{IV}}\}; Y). \tag{15}$$

It is not possible to evaluate the validity of the instrument Z on the basis of κ_i alone since it is an overall degree of confounding of the whole model. Even a positive κ_i could be consistent with a valid instrument if the confounding is due to $\tilde{U}$'s influence on $\mathbf{X}$ only (but not T_{IV}). In other words, there is no natural level which κ_i should be compared to. Replacing T_{IV} by a synthetic treatment variable T_s that is similar to T_{IV} though unconfounded solves this problem (cf. Panels A and B in Figure 2).

Specifically, we propose to generate a synthetic treatment variable T_s that has the same covariance structure to $\mathbf{X}$ as does T_{IV}, i.e. T_s satisfies

$$\text{Cov}(X_i, T_s) = \text{Cov}(X_i, T_{\text{IV}}) \; \forall i \in \{1, \dots, d\}.$$

See Algorithm 2 for the construction of T_s; we provide a detailed explanation of each step of that algorithm in Appendix B. On a high level, the idea is to generate a random variable W, regress out the variation in W that can be explained by $\mathbf{X}$, and then add parts of $\mathbf{X}$ back into W in a specific way that ensures that the resulting variable has the desired covariance structure w.r.t. $\mathbf{X}$. Then, we replace T_{IV} with that synthetic variable T_s and measure the degree of confounding in the resulting model:

$$\kappa_s := \kappa(\{\mathbf{X}, T_s\}; Y). \tag{16}$$

T_s is a synthetically generated variable that does not have a causal effect on Y and is, conditionally on $\mathbf{X}$, unconfounded while having the same covariance structure with $\mathbf{X}$ as T_{IV}. Thus, in terms of the definition of κ, cf. eq. (8), T_s is like T_{IV} except that it is not confounded (and that it has a causal effect of zero, a subtlety to which we return below). Intuitively, κ_s measures the counterfactual overall degree of confounding of the model that would be obtained if the instrument were valid and T_{IV} unconfounded (i.e. $b_{T_{\text{IV}}} = 0$). κ_s is the sought benchmark to which the actual degree of confounding κ_i can be compared to evaluate instrument validity.[4]

More formally, the following relation between the difference $\delta := \kappa_i - \kappa_s$ and instrument validity can be proven:

Theorem 1. *If the instrumental variable is valid, δ is not positive:*

$$IV \text{ valid} \Rightarrow \delta := \kappa_i - \kappa_s \leq 0. \tag{17}$$

4. The knockoff procedure by Candes et al. (2018) bears some similarity to Algorithm 2 as both are designed to generate variables that resemble their empirical counterpart in some form. We discuss how the two approaches relate and show that the proposed method is robust to using the knockoff variable procedure to generate T_s in Appendix K.

By contrapositive, this result implies the following corollary: If $\delta > 0$, the instrumental variable is invalid. The proof of these statements is found in Appendix A. Thus, the difference between the actual degree of confounding and the counterfactual degree of confounding, $\delta = \kappa_i - \kappa_s$, is informative about instrument validity: $\delta > 0$ implies instrument invalidity. Intuitively, if the instrument is invalid, instrumenting leads to a degree of confounding that is larger than the counterfactual degree of confounding. Loosely, δ can be interpreted as a "residual degree of confounding" of the instrumented model that is left after subtracting the confounding due to control covariates.

The corollary justifies evaluating the validity of Z on the basis of δ. If $\delta > 0$, we can deduce that the IV is invalid. The statement in Theorem 1 is a necessary but not sufficient condition for IV validity. We can only reject validity but can never reject invalidity. In other words, the instrument might still be invalid, even if $\delta \leq 0$.

It does not seem possible to provide a necessary *and* sufficient condition here. The reason for that lies in the way the synthetic variable T_s is constructed. T_s is unconfounded by construction ($b_{T_s} = 0$) and not causally related to Y (the true causal effect of T_s is equal to zero: $\tau_s = 0$). Precisely, κ_s measures the degree of confounding that would be obtained if the instrument were valid *and* $\tau = 0$. This drives a wedge between κ_i and κ_s even when the instrument is valid, cf. eq. (30). However, this does not affect the validity of Theorem 1, and therefore does not invalidate the test proposed here. If one could generate a synthetic variable with the same covariance structure as T_{IV} not only to $\mathbf{X}$ but also to Y, which would result in a estimated coefficient on that synthetic variable equal to the coefficient of T_{IV}, the wedge would close, the inequality in (29) would become an equality, and one could show that $\delta = 0$ if and only if the IV is valid. However, such a synthetic variable would be correlated with the unobserved confounder conditionally on $\mathbf{X}$ via its relation with Y, and thus would not induce a counterfactual degree of confounding suitable to compare κ_i with.

Since we do not observe the population quantities κ_s and κ_i, we rely on their estimates, denoted $\hat{\kappa}_s$ and $\hat{\kappa}_i$ respectively, to implement the test. We use the code provided by JS to estimate the κs. Similarly, to construct T_s we use sample covariances.

A formal derivation of the sampling error that underlie estimates of κ is not yet developed. To nevertheless incorporate uncertainty about $\hat{\kappa}_s$ and $\hat{\kappa}_i$, we calculate B estimates for κ_s and κ_i based on B bootstrap samples (lines 3 to 9 in Algorithm 1). For each bootstrap sample $b \in \{1, \ldots, B\}$ we calculate

$$\hat{\delta}_b = \hat{\kappa}_i - \hat{\kappa}_s \tag{18}$$

and the share of samples with $\delta_b \leq 0$,

$$\Delta_B = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\hat{\delta}_b \leq 0). \tag{19}$$

The resulting Δ_B can be interpreted as a pseudo-p-value for H_0. If the estimated actual degree of confounding ($\hat{\kappa}_i$) is larger than the estimated counterfactual degree of confounding ($\hat{\kappa}_s$) and Δ_B small, we have evidence for the IV being invalid, i.e. for rejecting H_0.

3.5 Behavior of pseudo-p-value and size of test

We call Δ_B a *pseudo-p*-value and not a p-value because it does not have a uniform distribution under H_0. However, we argue that it has a sub-uniform distribution under H_0. Consider the expression for δ under H_0 in eq. (30) in Appendix A. It implies that $\delta = \kappa_i - \kappa_s \approx \left\| \begin{pmatrix} \beta \\ \tau_s \end{pmatrix} \right\|^2 - \left\| \begin{pmatrix} \beta \\ \tau \end{pmatrix} \right\|^2 \leq 0$. As $d \to \infty$, the inequality becomes binding as the (infinitely many) β components of each vector outweigh the τ_s and τ (scalar) elements respectively. This implies $\delta = \kappa_i - \kappa_s = 0$ as $d \to \infty$ (under valid H_0). Within each bootstrap sample (see lines 3-9 of Algorithm 1), whether $\delta_b > 0$ or $\delta_b < 0$ is subject to chance. Therefore[5], the pseudo-p-value, which is the share of $\delta \leq 0$ across B bootstrap draws, will converge to 0.5 as $d \to \infty$. Thus, though the distribution of the pseudo-p-value Δ_B does not follow a uniform distribution, the cumulative distribution function F of Δ_B has the following property: $F(\Delta_B \leq t) \leq t$. Namely, it is sub-uniform. This implies that the false positive rate of the test lies below the nominal size of the test.

With finite d, nonzero causal effect ($\tau \neq 0$) and under H_0, δ is smaller than 0 since $\tau_s = 0$ by construction. In other words, δ is strictly negative under H_0 when $\tau \neq 0$. Therefore, Δ_B tends to be larger than 0.5. As a consequence, the empirical distribution of Δ_B has a left skew. Figure 6 shows histograms of the pseudo-p-value for simulated data with valid instruments, which underscore this theoretical point. Although we cannot guarantee that the test has asymptotically exact size under H_0, the behavior of the pseudo-p-values implies that the size of the test lies below its nominal size α. In other words, we can guarantee size control. The lack of an exact size guarantee would be problematic if it came at the cost of low power. While the AUC curves in Figure 7 do not directly show the test's power (as they show sensitivity-specificity trade-offs), their high levels (above 0.8 as the endogeneity of the instrument increases) indicate that the procedure does not suffer from low power.

It is worthwhile stressing that size control, not exact size guarantee, is typically achieved in the nonparametric testing literature (see e.g. Breunig and Chen, 2020; Fang and Seo, 2021; Li et al., 2022). While the proposed test assumes rotation-invariant priors for the model parameters and is therefore not non-parametric, this observation stresses that size control is more important than an exact size guarantee.

3.6 Inherited assumptions from Janzing and Schölkopf (2018a)

The method proposed by JS relies on a high-dimensional set of covariates to estimate a degree of confounding. Since this method is a central part of the proposed test, we inherit that reliance on high-dimensional set of $\mathbf{X}$. In other words, we require a sufficiently large set of control variables in addition to the treatment variable T for the test to work. In practice, this reliance on additional control variables $\mathbf{X}$ is not limiting as it is unlikely *not* to have additional control variables. Moreover, both the empirical applications as well as the Monte Carlo study show that already around five covariates work well in practice.

In addition, JS make idealized assumptions about the generating process of the linear model they study to show how to estimate a degree of confounding. More specifically, they require the structural parameters to be rotation-invariant. We describe this assumption in

5. Technically, this argument relies on $B \to \infty$, which we disregard here.

Data: sample of the outcome variable, control covariates, treatment indicator, and instrumental variable $\mathcal{D} = \{Y_i, \mathbf{X}_i, T_i, Z_i\}_{i=1}^n$

Input: data $\mathcal{D}$, threshold value α, number of bootstraps B

Output: pseudo-p-value and rejection decision $\psi(\alpha)$ for the hypothesis $H_0:$ Z is a valid instrument

Normalize data such that all variables have equal means and equal variance, e.g. a mean of zero and a variance of one.

Implement two-stage least squares IV approach: regress T on $\{\mathbf{X}, Z\}$, compute the fitted values and call them T_{IV}

for $b = 1$ **to** B **do**

  Draw a bootstrap sample $\mathcal{D}_b$ of size n with replacement

  Estimate $\kappa_i := \kappa(\{\mathbf{X}, T_{\text{IV}}\}; Y)$ based on $\mathcal{D}_b$ following JS, call estimate $\hat{\kappa}_i$ (See Appendix H for a description how to estimate κ.)

  Generate synthetic variable T_s based on $\mathcal{D}_b$ by following Algorithm 2

  Estimate $\kappa_s := \kappa(\{\mathbf{X}, T_s\}; Y)$ based on $\mathcal{D}_b$ following JS, call estimate $\hat{\kappa}_s$

  Calculate $\hat{\delta}_b = \hat{\kappa}_i - \hat{\kappa}_s$

end

Calculate the pseudo-p-value

$$\Delta_B = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\hat{\delta}_b \leq 0)$$

Decide whether to reject H_0: $\psi(\alpha) = \mathbb{1}(\Delta_B \leq \alpha)$

Algorithm 1: Test for instrument validity

detail in Appendix H, see in particular Assumption 5. While that assumption of rotation-invariant structural parameters is required for the theoretical derivation of the method to estimate κ, JS write that "[t]here is some hope that empirical data show similar concentration of measure phenomena although our model assumptions are probably significantly violated" (p. 24). In our Monte Carlo studies (see Section 5) we generate data that does not have rotation-invariant priors and, yet, the proposed method performs well. In addition, it is unlikely that the data used in the empirical applications in Section 6 is generated from a process with rotation-invariant priors. Yet, the proposed method performs well. This substantiates the claim made by JS. Note that for the derivation of Theorem 1, rotation-invariant priors need not be assumed.

3.7 Discussion of constant treatment effects assumption

Much of the recent literature on inference in IV models and testing IV assumptions relies on heterogeneous treatment effects models, where the parameter of interest is the Local Average Treatment Effect (LATE, i.e. the treatment effect for a subpopulation called the 'compliers', i.e. those members of the population whose treatment status depends on the value of the instrument, see the seminal paper by Imbens and Angrist, 1994). Unlike this strand of the literature, we focus on a linear model with additively separable error term

Data: $n \times d$ matrix of covariates $\mathbf{X}$ with full column rank,
$n \times 1$ vector of instrumented treatment T_{IV}
Output: a random variable T_s with $\text{Cov}(X_i, T_s) = \text{Cov}(X_i, T_{\text{IV}}) \, \forall i \in \{1, \ldots, d\}$

1 Define the vector $\rho := \begin{pmatrix} \text{Cov}(X_1, T_{\text{IV}}) & \ldots & \text{Cov}(X_d, T_{\text{IV}}) \end{pmatrix}^\top$
2 Draw n samples of a standard Gaussian $\mathcal{N}(0,1)$, and collect those samples in a $n \times 1$ vector W.
3 Regress W on $\mathbf{X}$ and compute residuals: $\eta := W - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top W$
4 Compute the singular value decomposition of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^\top$ where the diagonal elements $\{\sigma_j\}_{j=1}^d$ of $\mathbf{S}$ are the singular values of $\mathbf{X}$, $\mathbf{U}$ contains the left-singular vectors, $\mathbf{V}$ contains the right-singular vectors
5 Compute $\mathbf{X}_{\text{dual}} := (n-1) \times \mathbf{U} \times diag(1/\sigma_j) \times \mathbf{V}^\top$
6 Compute $s := \sqrt{\frac{1 - \rho^\top \times \text{Cov}(\mathbf{X}_{\text{dual}}) \times \rho}{\text{Cov}(\eta)}}$
7 Compute $T_s := \mathbf{X}_{\text{dual}} \times \rho + s \times \eta$

Algorithm 2: Generate synthetic T_s

and a constant treatment effect τ in this paper. The main reason for that is the following. Since the proposed test relies on the method to estimate a degree of confounding in *linear models with additively separable error term* developed in JS, we are restricted to work in the same model class. Extending JS to nonlinear models, is an interesting avenue for future research. Such an extension could make the idea of constructing a counterfactual degree of confounding under instrument validity amenable to heterogeneous treatment effects models in the future.

Although it is common to use heterogeneous treatment effects models to estimate LATE, the focus on LATE as the estimand of interest is controversial. For example, Deaton (2010) argues that "we are unlikely to learn anything about the processes at work" (p. 490) if we are unwilling to make structural assumptions, such as the ones in eqs. (4)-(5), that allow us to estimate structural parameters, i.e. average treatment effects (ATE). Deaton urges researchers to focus on describing *mechanisms* (what he calls 'processes') as those are more useful to decision-makers than LATE. This is consistent with the use of the Principle of Independent *Mechanisms* to test the validity of IV assumptions.

In the linear, constant treatment effects model, Heckman and Vytlacil (2005) show that additional testable implications arise. Specifically, they show how to exploit distributional information that goes beyond the second-order moments to test IV validity. One contribution of the present article is that we can provide a test for IV validity without relying on such higher-order moments, but rather by exploiting implications of the Principle of Independent Mechanisms. As such, our approach shows that other sources of information can be used to assess IV validity. We want to stress, however, that our testable implication is not sharp in the sense of Balke and Pearl (1997).

3.8 Generalization to models with high-dimensional confounders

Though the model in eqs. (4)-(5) has a one-dimensional confounder U, our testing framework can easily be extended to models with higher-dimensional confounders. Specifically,

Janzing and Schölkopf (2018b) (JSb) show how to estimate a degree of confounding κ in models with high-dimensional confounders. Though the methodological approach between these two papers (JS and JSb) differs, the definition of the degree of confounding κ that both approaches estimate is exactly the same.

Our test procedure carries over seamlessly to cases with high-dimensional confounders because we can swap the method to estimate κ in lines 5 and 7 of Algorithm 1 from JS (one-dimensional confounder) to JSb (high-dimensional confounder). We provide a summary of the main arguments of JSb in Appendix I and reproduce all simulations results as well as empirical applications using the method laid out in JSb in Appendix J. These results are similar to those discussed in the main text.[6]

4. The Principle of Independent Mechanisms and Generic Orientation

This section describes and illustrates the idea of the method to estimate a degree of confounding in multivariate linear models proposed by Janzing and Schölkopf (2018a). The main arguments of JS are reproduced formally in Appendix H. In particular, Assumption 5 about rotation-invariant structural parameters is specified formally.

The Principle of Independent Mechanisms (PIM) underlies many contributions to causal inference from the machine learning community (for an overview see Schölkopf et al., 2021). It also serves as the basis for the test proposed in this paper. The notion goes back to pioneering econometricians Haavelmo and Frisch, who identified the search for and analysis of independent mechanisms as the ultimate goal of econometrics (though using slightly different terminology, calling them "autonomous", see Frisch et al., 1938). Despite considering it an important guiding principle, they did not employ the notion of independent mechanisms as an empirical identification technique as such. In fact, Frisch and Haavelmo argued that the independent nature of mechanisms cannot be identified from observational data but must be motivated by (economic) theory (or controlled experiments). Some advances towards its use as an identification tool for studies based on observational data have been achieved. The proposal by Janzing and Schölkopf (2018a) to estimate the degree of confounding in multivariate linear models, which is motivated by the notion of independent mechanisms, is an example of this progress.

To illustrate the idea, consider a set of random variables $\{V_1, \ldots, V_n\}$ whose causal relations can be represented in a directed acyclic graph (DAG) and an accompanying structural equation model (Pearl, 2009). The joint probability distribution that is consistent with the

6. A tangential comment is in order here. As described in Appendices H and I, Janzing and Schölkopf provide two entirely independent ways of estimating the degree of confounding κ for multivariate linear models with one-dimensional and high-dimensional confounders, respectively. Each approach makes idealized (and very different) modeling assumptions. Still, both approaches lead to very similar results when used in the present setting of evaluating instrument validity. This is remarkable and testifies to the ingenuity of Janzing and Schölkopf's work. It also lends credibility to the testing framework proposed here; specifically, the idealized modeling assumptions in JS and JSb are not what is driving the results presented in this paper (keep in mind also that the simulation setting we use already departs from the idealized assumptions in both JS and JSb).

causal structure given in the DAG can be factorized as

$$P(V_1, \ldots, V_n) = \prod_{j=1}^{n} P(V_j | Pa(V_j)) \tag{20}$$

where $Pa(V_j)$, the *parents* of V_j, denotes the set of random variables that directly cause V_j. Naturally, there are many other types of factorizations of the joint distribution:

$$P(V_1, \ldots, V_n) = \prod_{j=1}^{n} P(V_j | V_{j+1}, \ldots, V_n). \tag{21}$$

However, only the factorization in eq. (20) is a description of the data generating process implicit in the DAG, which represents the causal generating mechanisms of the data: first "nature" generates data for the parental nodes, these feed into the descendant nodes ("children"), etc. The conditionals in eq. (20) represent causal mechanisms that translate causes (or parents, $Pa(V_j)$) into their effects ("children", V_j). Causes that do not have parents in the model under investigation appear as marginal distributions in this formulation. Using algorithmic information theory, Janzing and Schölkopf (2010) and Lemeire and Janzing (2013) show that the conditionals on the right-hand-side are algorithmically independent of each other if the DAG represents the causal structure. Intuitively, changing one mechanism, e.g. by intervening to set the corresponding child variable to a specific value, does not change any other mechanism. In this sense each of the mechanisms operates independently of the others. Since the formal deduction of the mechanism's algorithmic independence relies on the theoretical notion of Kolmogorov complexity, which cannot be estimated, it is not obvious how to conceive of the independence of mechanisms in practice. Thus, the algorithmic independence of mechanisms amounts less to a precise recipe for uncovering autonomous relations in observational data than to a rigorous guiding principle to design algorithms that do.

To make the notion of 'independent mechanisms' practically relevant, what precisely is meant by 'independence' must be defined in a way that allows data-driven quantification. Janzing and Schölkopf (2018a) propose such a feasible interpretation of the Principle of Independent Mechanisms. Moreover, they show a way to measure the degree of violation of PIM in observational data. This degree of violation is a measure of confounding in multivariate linear models. We spend the rest of this section illustrating their notion of independence and how they can infer a measure of confounding. This is not an exhaustive discussion. The technical details are provided in Appendix H, which reproduces the arguments in JS.

To illustrate the proposal by Janzing and Schölkopf (2018a), consider two versions of a simple multivariate linear model $Y = \mathbf{X}\beta + \varepsilon$. First, in the unconfounded model, $\varepsilon \perp\!\!\!\perp \mathbf{X}$. The multidimensional $\mathbf{X}$ is causing Y and the least-squares estimate of β is unbiased. Second, in the confounded model, $\varepsilon \not\perp\!\!\!\perp \mathbf{X}$. The multidimensional $\mathbf{X}$ is still causing Y but now due to the dependence of ε with $\mathbf{X}$, the least-squares estimate of β is biased. How to infer from observational data whether the least-squares estimate of β is biased?

To see how Janzing and Schölkopf (2018a) answer that question, note first that the true β is the crucial parameter representing the 'mechanism' that translates the causes $\mathbf{X}$

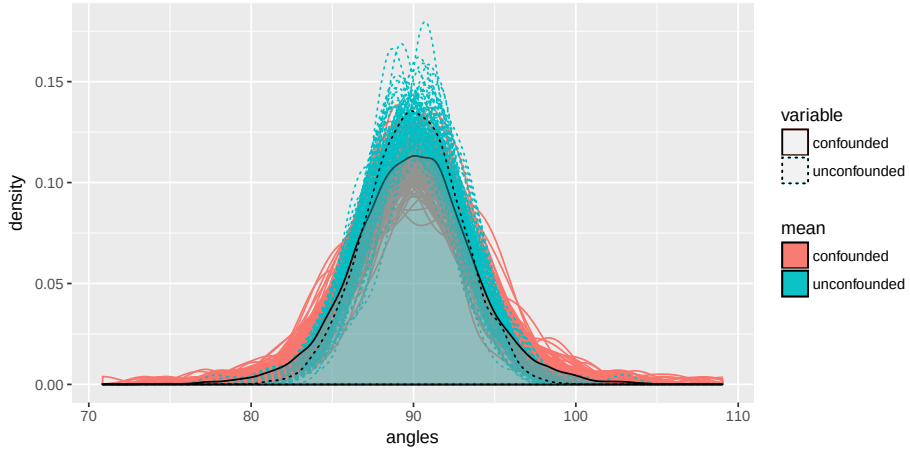

Figure 4: *Illustration of genericity of causal parameter vectors. This figure shows density plots of the angles between the least-squares parameter vector of both confounded and unconfounded models with each of the d eigenvectors of the covariance matrix of the covariates. In the unconfounded model, the least-squares parameter vector should lie in generic orientation with respect to (the eigenspace spanned by the) eigenvectors of the covariance matrix of the covariates. Genericity of two vectors can be understood as their dot product being zero. As expected, therefore, the distribution of angles in the unconfounded case clusters around 90 degrees. Crucially, in the confounded case, the distribution of angles is considerably wider. A trace of confounding is thus reflected in the less generic angles of the confounded parameter vector with respect to the eigenvectors; their distribution is characterized by a more frequent divergence from the generic angle of 90 degrees. This illustrates the type of confounding signal that JS leverage in their methodology. The figure shows angle distributions for 100 simulation runs with* $d = 100$*, and* $n = 50000$*, the respective means are depicted with black lines, solid for the confounded and dashed for the unconfounded case. Details on the simulation setting is found in Appendix G.*

into effect Y. The causes, in turn, are represented by the covariance matrix of the right-hand-side variables $\Sigma_{\mathbf{XX}}$. What the PIM implies on an intuitive level is that the *mechanism* translating causes into effect, represented by the true parameter vector, and the *input to the mechanism* or *causes*, represented by $\Sigma_{\mathbf{XX}}$, should be 'independent'. JS make the concept of 'independence' estimable by arguing that, if PIM is fulfilled, the true parameter vector should lie in generic orientation with respect to the eigenspace spanned by the eigenvectors of the covariates' covariance matrix, $\Sigma_{\mathbf{XX}}$. In technical terms, such genericity is defined by the equivalence of two spectral measures: the spectral measure of $\Sigma_{\mathbf{XX}}$ induced by the true parameter vector (which results from weighting the eigenvalues of $\Sigma_{\mathbf{XX}}$ by that true parameter vector) should be equal to the (unweighted) tracial spectral measure of $\Sigma_{\mathbf{XX}}$.

We now provide a graphical illustration of the traces that a violation of PIM leaves in purely observational data. We simulate data from a confounded and an unconfounded model, then estimate the parameter vector by least-squares in both cases (see Appendix G for details on the simulation). In the unconfounded case, the estimated parameter vector

represents genuine causes and is not biased due to unobserved confounding. Following JS, that true parameter vector should lie in generic orientation with respect to the eigenvectors of the covariance matrix. Two vectors lie in generic orientation with respect to each other if their dot product is zero (or the angle they span is ninety degrees). At first glance orthogonality seems like a specific, not generic, relation between any two vectors. However, it is important to note that such genericity is a high-dimensional phenomenon: the angle between two randomly drawn vectors approaches ninety degrees as the their dimensionality increases (see e.g. Gorban and Tyukin, 2018). This is also why the asymptotic results in JS rely on the dimensionality of the covariate space going to infinity. Intuitively, two generic vectors do not share any information since they are pointing in two orthogonal directions.

Therefore, we compute the angle between the estimated parameter vector and each of the eigenvectors of the covariance matrix of the covariates for both the confounded and unconfounded setting and plot their distribution.[7] For both settings, we simulate data for $d = 100$ dimensions and $n = 50,000$ observations resulting in 100 calculated inner products. Then, we plot the resulting distribution of angles between d eigenvectors and the least-squares estimate $\hat{\beta}$. Figure 4 plots these distributions for 100 draws of the data. Crucially, one can see that the distribution of angles is more widespread for the confounded setting. Consequently, in the presence of confounding the estimated parameter vector lies in a less generic direction with respect to the eigenvectors of the covariance matrix. This deviation from genericity is what Janzing and Schölkopf (2018a) exploit to measure the degree of confounding.

5. Monte Carlo Simulation

To see how the proposed instrument validity test performs, we run Monte Carlo studies. In the main body of the paper, we present the simulation to study violations of the exclusion restriction. The simulation to study violations of the exchangeability assumption are relegated to the Appendix F since the results and conclusions are similar. We are distinguishing between two simulation settings (violations of exclusion restriction and exchangeability assumption) to show that the proposed test can detect violations of either assumption, although it is not able to distinguish which assumption is violated. Note that we are simulating data for a setting with a binary treatment variable to ensure better comparability to Kitagawa (2015), although the theoretical development of the test requires a continuous treatment variable.

Data that simulates a violation of the exclusion restriction is generated according to Algorithm 3. Simulations are implemented for each combination of the following parameters: number of observations: $n \in \{100, 500, 1000\}$, number of covariates: $d \in \{5, 10\}$, degree of violation of the exclusion restriction: $\omega_1 \in \{0, 0.1, 0.2, 0.4, 0.5\}$, degree of the relevance of the instrument: $\omega_2 \in \{0.3, 0.6\}$. Moreover, the following parameters are fixed: number of bootstrap samples $B = 100$, number of Monte Carlo draws $M = 200$.

7. For the purpose of illustration, we depart slightly from JS here. We compute the genericity of the estimated parameter vector for every eigenvector *in isolation*. However, JS postulate a generic orientation with respect to the eigenspace spanned by the collection of eigenvectors. In other words, they jointly consider the whole set of eigenvector-eigenvalue pairs. Technically, they consider the distribution of eigenvalues weighted by the estimated parameter vector. See Appendix H.

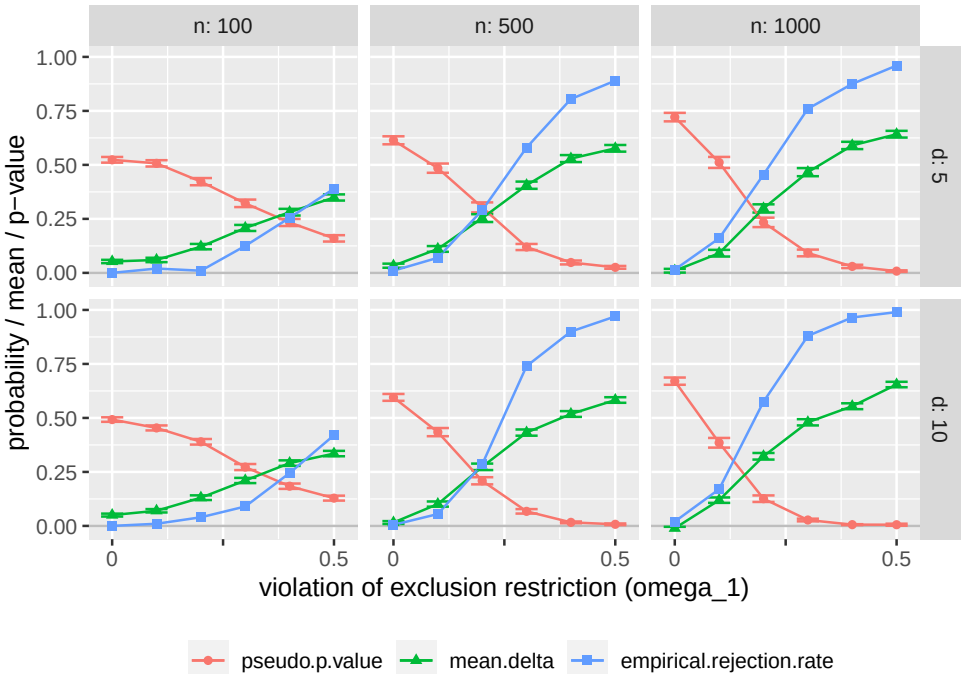

Figure 5: *Simulation results: pseudo-p-values, δ_B, and empirical rejection rate as a function of ω_1. This figure shows averages over all M Monte Carlo draws of the p-value, δ_B, and the empirical rejection probability (based on the p-value with threshold parameter $\alpha = 0.05$) as a function of the degree of violation of the exclusion restriction (ω_1), by number of covariates d and number of observations n. $\omega_2 = 0.3$. δ_B rises sharply with ω_1, the pseudo-p-value decreases as ω_1 increases. Consequently, the empirical rejection probabilities increase as ω_1 increases indicating that, if the degree of condounding is sufficiently high, the test rejects the null of instrument validity in all Monte Carlo draws.*

The following key statistics are reported: the pseudo-p-value, the average difference between κ_i and κ_s over all boostrap draws, $\delta_B = \frac{1}{B}\sum_{b=1}^{B}(\hat{\delta}_b)$, as well as the empirical rejection rate for a threshold value of $\alpha = 0.05$, i.e. we reject when $\Delta_B < \alpha$.

Figure 5 shows the evolution of the average over 200 Monte Carlo runs of pseudo-p-value and δ_B as a function of the degree of violation of the exclusion restriction (ω_1). Both measures are increasing with ω_1, which shows that they are sensitive to the confoundedness of T_{IV}. The empirical rejection rate based on the pseudo-p-value with $\alpha = 0.05$ increases as a function of ω_1. The null hypothesis of instrument validity is rejected increasingly often as ω_1 is rising. Generally, both a larger d and a larger n improve the performance of the test; however, given d, increasing n improves performance by more than increasing d given n. Considering that the asymptotic results in JS require n and $d \to \infty$, we show Monte Carlo results for very small d. Still, the test works well.

Figure 6 shows empirical pseudo-p-value distributions under H_0 for various combinations of n and d. As a benchmark, the horizontal line shows how the histograms would look like if the pseudo-p-value Δ_B had a uniform distribution. The theoretical claims about the

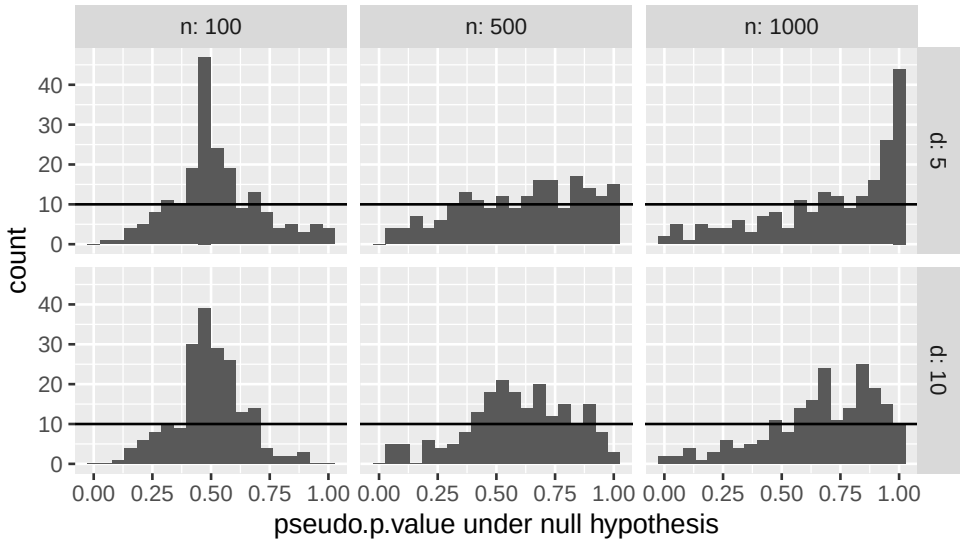

Figure 6: *Simulation results: distribution of pseudo-p-value under H_0. This Figure shows histograms of pseudo-*p*-values under H_0 : instrument validity, i.e. when $\omega_1 = 0$ for combinations of number of observations and number of covariates. $\omega_2 = 0.3$. The source of confounding is a violation of the exclusion restriction. The horizontal bar indicates the corresponding histogram for a uniform distribution. Though the pseudo-*p*-values are not uniformly distributed, they follow a sub-uniform distribution, which implies that the false positive rate lies below the nominal size of the test.*

distribution of Δ_B under H_0 from Section 3.4 are substantiated empirically here: while Δ_B does not have a uniform distribution, it has a sub-uniform distribution. This implies that the empirical size of the test lies below the nominal size α.

To evaluate the trade-off between making type I and type II errors we calculate the area under the ROC curve (AUC) and plot it as a function of ω_1 in Figure 7. A type I error is committed when the test rejects the validity of the instrument (H_0) although, in fact, the instrument is valid. The AUC levels are increasing rapidly as ω_1 increases and reach values above 0.9 when d and n are large. It is noteworthy that the AUC levels tend to be larger for a lower value of the degree of relevance of Z (ω_2). As ω_2 increases Z contains less and less variation in addition to that in T that can be leveraged in the IV implementation or in the validity test. In the extreme, Z and T collapse to one variable and the instrumented T does not contain any different information than T. In other words, the instrument cannot extract the experimental variation of T when ω_2 is too large. Nevertheless, even for large ω_2, the proposed test performs well with AUC levels ranging from 0.6 (low degree of endogeneity of instrument) to 0.9 (high degree of endogeneity).

We compare the performance of our instrument validity test to the one proposed by Kitagawa (2015). Note that these two tests rely on two entirely different approaches: the former on the genericity of estimated parameter vectors with respect to the covariance

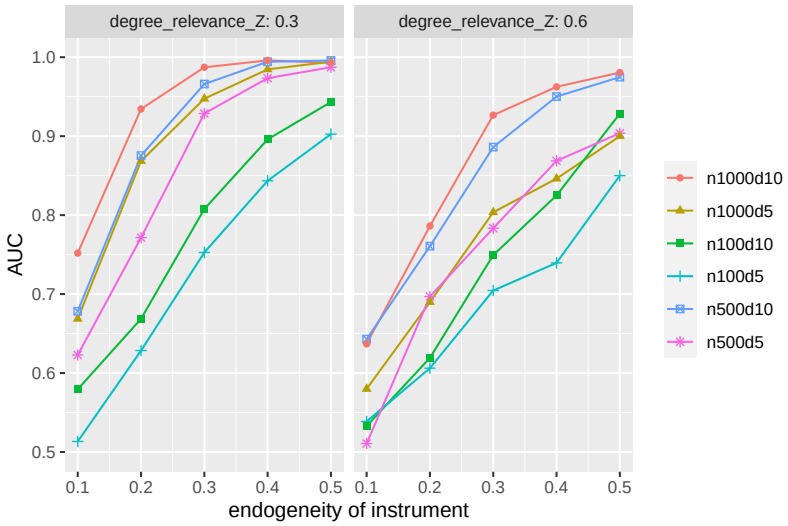

Figure 7: *Simulation results: AUC curves. This Figure shows the area under the ROC curve (AUC) as a function of the degree of violation of the exclusion restriction (ω_1), for various combinations of number of covariates, d, and number of observations, n, by instrument relevance degree (ω_2, horizontal). The underlying test statistic is the pseudo-p-value. The test achieves AUC levels above 0.9 for large ω_1, n, and d. A increase in ω_2 implies an increase in the correlation between Z and T. As this correlation becomes larger there is fewer variation in Z to extract experimental variation from T, resulting in decreased performance of the algorithm .*

matrix of independent variables, the latter on checking whether distributions of Y for four subgroups identified by the interaction of two binary variables, T and Z, intersect.

Figure 8 shows comparisons of AUC levels for the test proposed in this paper and the one proposed by Kitagawa (2015). The AUC levels for our approach generally lie above those corresponding to Kitagawa's approach. In the linear model studied here, our approach outperforms Kitagawa's especially for low levels of ω_1. However, Kitagawa's approach is also applicable in nonparametric models and heterogeneous treatment effects models, where our approach does not apply. Moreover, our approach relies on additional structural assumptions, in the form of the Principle of Independent Mechanisms, which Kitagawa does not need (see Section 3.6). On the other hand, our approach is not restricted to binary treatment and binary instrument IV models. Note that the constant treatment effects model that we analyze in this paper is a special case of the model studied by Kitagawa.

An important limitation of the algorithm proposed by JS is that the estimated κ is, in theory, not robust to rescaling of the data as this introduces a dependence between the covariance matrix of the covariates and the parameter vector. For instance, consider income as measured in thousands of USD. Its rescaling by logarithms changes both the covariance structure of independent variables and the parameter vector. The authors acknowledge this, yet claim and show in simulations that the estimated κ is robust to rescaling of the data

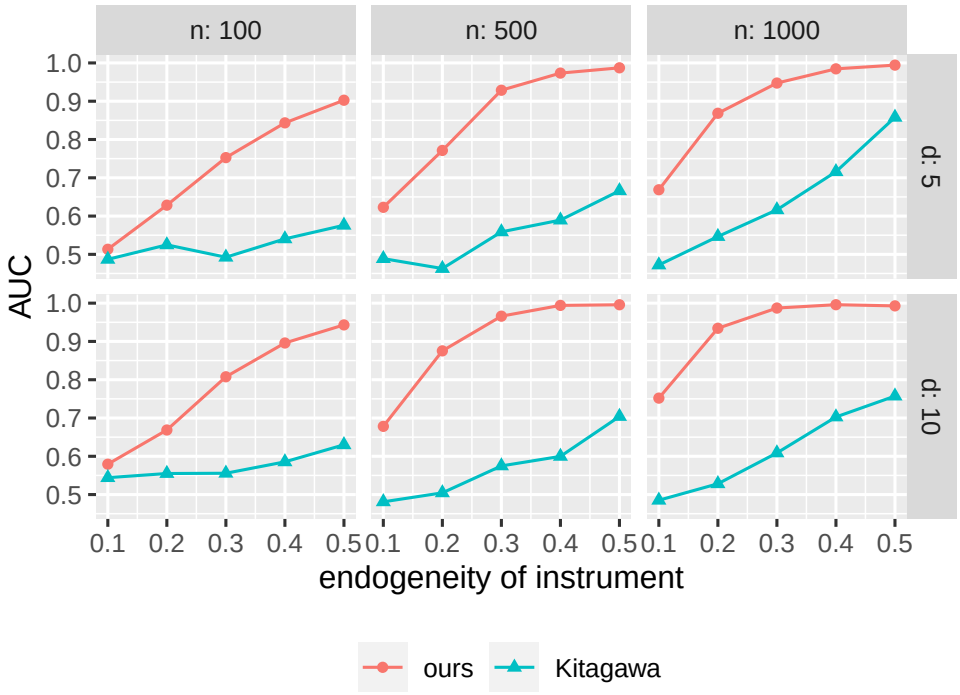

Figure 8: *Simulation results: Comparison to Kitagawa. This Figure shows the Area Under the ROC curve (AUC) for the test proposed by Kitagawa (2015) and ours for combinations of number of observations and number of covariates as a function of the degree of violation of the exclusion restriction* (ω_1)*.* $\omega_2 = 0.3$*. Note that Kitagawa's test is applicable more widely in non-parametric models and not only in the linear setting studied here.*

in practice.[8] The proposed test relies on a comparison of *two* κs, which is useful beyond the fact that such a comparison allows focusing on the bias of *one* covariate: Both κs are influenced by transformations in the same way, which one can therefore expect to leave the sign of their differences, i.e. δ, unaffected. In Appendix D we document the robustness of the proposed algorithm to typical data transformations: the observed AUC levels are insensitive to the implemented transformations of the data and the pseudo-p-values of the validity test on untransformed and transformed data show a correlation coefficient around 0.9.

In Simulation Regime 2 we analyze whether the algorithm can also detect an invalid instrument when its invalidity stems from the violation of the exchangeability assumption.

8. An interesting insight in this context is due to Holmes and Caiola (2018). A given regression techniques should fulfill certain properties to be useful. Two such properties are scale invariance (it should not matter whether data is measured in centimeters or inches) and rotational invariance (it should not matter 'from which angle you are looking at the data'). As an example, ordinary least-squares is scale-invariant but not rotationally invariant; Principal Component Analysis is rotationally invariant but not scale-invariant. Holmes and Caiola derive the incompatibility of these two criteria. For this reason, it might not seem surprising that the JS methodology, which relies on some limited type of rotational invariance, is not scale-invariant. Note that JS assume rotational invariance of the prior on the structural parameter vectors; they do not assume rotational invariance of the model itself.

The results are presented in Figures 12 and 13 in Appendix F. The performance of the test is similar. We provide additional simulation results for the case when JSb is used to estimate κ_i and κ_s in Appendix J. The results are robust to this choice of estimation method.

Input: number of observations n, number of covariates d, variance of the structural errors σ^2;
two parameters specifying relation between instrument, treatment and unobserved confounder: ω_1: degree of violation of exclusion restriction, ω_2: the relevance of the instrument Z

Output: a simulated data set of n observations of outcome variable, covariates, treatment variable, instrument $\mathcal{D} = \{Y, \mathbf{X}, T, Z\}_{i=1}^{n}$

1 **Generation of structural errors**: ε_Y and ε_T, are drawn from

$$\begin{pmatrix} \varepsilon_Y \\ \varepsilon_T \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right) \tag{22}$$

2 **Generation of instrument** Z: Let $Z \sim \text{Bernoulli}(0.5)$.

3 **Generation of covariates** $\mathbf{X}$: Draw d eigenvalues from a uniform distribution $\lambda_i \sim \mathcal{U}(0.5, 1.5)$ which populate the diagonal of a $d \times d$ matrix Λ. Then draw a random orthonormal matrix $\mathbf{O}$ of dimension (d), set $\Sigma = \mathbf{O}\Lambda\mathbf{O}^\top$ and draw $\mathbf{X}_{\text{temp}}$ from a multivariate normal distribution $\mathbf{X}_{\text{temp}} \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

4 Draw a random (d)-dimensional vector from a normal distribution: $\beta_{c,temp} \sim \mathcal{N}(0, 1)$ and, to keep the variance of Y comparable for various d, normalize $\beta_c = \beta_{c,temp} / \left\|\beta_{c,temp}\right\|$. With these ingredients set

$$\mathbf{X} = \mathbf{X}_{\text{temp}} + \varepsilon_Y \beta_c^\top. \tag{23}$$

5 **Generation of treatment** T: To induce dependence of the treatment on the set of covariates, first draw the d-dimensional vector $\beta_{T,temp}$ populated with draws from a $\mathcal{N}(0, 1)$, $\beta_{T,temp} \sim \mathcal{N}(0, 1)$ and set $\beta_T = \left(\beta_{T,temp}\right) / \left\|\left(\beta_{T,temp}\right)\right\|$ to keep the relative influence of $\mathbf{X}$ on T comparable for various d.

6 Generate T, as $T = \mathbb{1}\left(\mathbf{X}\beta_T^\top + \omega_2 Z + \varepsilon_T > T'\right)$ where T' is the mean of $\mathbf{X}\beta_T^\top + \varepsilon_T$ and $\mathbb{1}$ is the indicator function.

7 **Generation of the outcome variable** Y: First generate a random d-dimensional vector $\beta_{\text{temp}} \sim \mathcal{N}(0, 1)$. To keep the variance of Y comparable for various d, set $\beta = \left(\beta_{\text{temp}}\right) / \left\|\left(\beta_{\text{temp}}\right)\right\|$. The true causal effect of the treatment variable is set to $\tau = 1$. Finally, generate outcome Y as

$$Y = \mathbf{X}\beta^\top + \omega_1 Z + \tau T + \varepsilon_Y. \tag{24}$$

Normalization of data To keep the binary nature of T, we normalize the data to have equal variance and equal mean as T (as opposed to normalizing all variables to have zero mean and variance equal to one).

Algorithm 3: Simulation of Violation of Exclusion Restriction

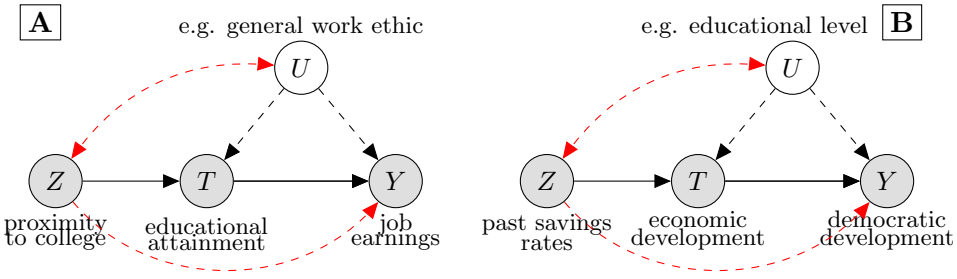

Figure 9: *Two examples of IV models Panel A: Estimating the causal effect of educational attainment on job earnings (Card, 1995). Panel B: Estimating the causal effect of economic development on democratic development (Acemoglu et al., 2008).*

6. Two empirical applications

We apply the proposed method to two empirical IV studies, see Figure 9. First we use data from Card (1995) to test the validity of {proximity to college} as an instrument for {educational attainment} in an effort to estimate the causal effect on {earnings}, see Section 6.1. Second, we apply the test to evaluate the validity of {past saving rates} as an instrument for {economic development} in a study by Acemoglu et al. (2008) that attempts to understand its causal effect on {democratic development}, see Section 6.2. In both applications, we discuss how PIM can be interpreted in the context at hand.

We provide results for both applications also when JSb is used to estimate κ_i and κ_s in Appendix J.

6.1 Empirical application to Card (1995)

Estimating financial returns to education is a long-standing problem in labor economics. A specific question is what causal effect does spending an additional year at college have on subsequent job earnings. The relation between educational attainment and subsequent job earnings is marred by many unobserved confounding variables such as general ability or parental socio-economic status. Since experimentally controlling levels of education to estimate the causal effect of education is not feasible, economists have used instrumental variable designs to estimate the causal effect of education on earnings. For example, Card (1995) proposes the proximity of a family's residence to a four-year college as an instrument of the children's educational attainment to estimate financial returns to education. Panel A in Figure 9 depicts the returns to education study.

Card himself casts doubt on the validity of college proximity as an instrument as there might be factors such as family preferences or local labor market conditions that are related to both the proximity to a college and the outcome variable: families might move closer to colleges *because* they expect higher earnings for their offspring in vibrant labor markets found in proximity to colleges, which would constitute a violation of the exclusion restriction. Card uses a sample of roughly 3,500 individuals from the National Longitudinal Surveys of Youth (NLSY, Cooksey, 2018), which collects longitudinal data on a cohort of baby boomers on cognitive and socio-emotional development as well as socio-economic status, educational attainment and subsequent job earnings. Card argues that while the proposed instrument

is not valid in the whole sample, it is likely valid after controlling for a set of socio-economic variables $\{S\}$:= {ethnicity dummy, father's educational level, living in South dummy for 1966 and 1976, urban residence dummy for 1966 and 1976}. We binarize the treatment indicator to equal 1 if years of education is greater than 16 years, i.e. the treatment can be considered as getting a college degree.

Succinctly, though {college proximity} is not a valid instrument unconditionally, it is a valid instrument conditionally on {S}. This is Card's claim that we will corroborate by using the test for instrument validity proposed in this paper. We run our test three times: first, we include the full set of 29 covariates which include, beyond $\{S\}$, diverse information on estimated IQ levels, the Knowledge of the World score, availability of a library card in the household head's childhood home, marital status, labor market experience, etc. Call this set of additional covariates $\{R\}$. Second, we include only variables $\{S\}$. In these first two cases, we expect the test not to reject instrument validity since we are controlling for those variables $\{S\}$ that render the instrument valid according to Card. Third, we include only variables $\{R\}$ and exclude variables $\{S\}$. In the third run, if Card's argument holds, one would expect the test to reject the null hypothesis of instrument validity since the crucial set of covariates {S} is omitted. Table 1 reports the pseudo-p-value for each of the three sets of covariates. The results show that, indeed, the test does not reject the null of instrument validity if set $\{S\}$ is controlled for. On the contrary, once $\{S\}$ is left out of the set of covariates, the test rejects instrument validity. This corroborates Card's argument. Furthermore, these results show that the proposed test is able to detect validity of the instrument solely based on the spectra of the covariates induced by the estimated parameter vectors. In Appendix C, we report results of the proposed test for settings where all possible subsets of $\{S\}$ are included as covariates.

This discussion of Card's argument illustrates how conditioning on a (possibly large) set of covariates is often necessary to render an instrument valid. The test proposed in this paper can naturally evaluate conditional instrument validity. Accounting for additional variables simply amounts to including more covariates in the IV regression in Step 2 of Algorithm 1.

6.2 Empirical application to Acemoglu et al. (2008)

A positive correlation between measures of democracy and per capita income is an empirical regularity (Acemoglu et al., 2008). Many OECD countries that score high on democracy measures also have high per capita income levels. Vice versa, many non-democracies, e.g. in Subsaharan Africa and Southeast Asia, have relatively low levels of per capita income. Though this empirical pattern is often explained by hypothesizing that higher income *causes* political institutions to become more democratic (see e.g. Huntington, 1991), it is difficult to assess the credibility of such claims. Third factors might cause a country to embark on a democratic development path as well as increase its per capita income whithout a direct causal effect between the two. To complicate matters, the presumed causal relation might go into the other direction, i.e. democratic institutions might cause higher future growth (see e.g. Acemoglu et al., 2019). It is not our ambition to solve this question here; rather, we want provide an additional application of our method to real-world data.

	test results for different sets of covariates		
	$\{R, S\}$	$\{S\}$	$\{R\}$
pseudo-p-value	0.18	0.29	0.00
no. of covariates	29	6	23
no. of observations	3612	3612	3612

Table 1: *Results of empirical application to Card (1995). This Table shows results of the empirical application, based on the data used by Card (1995). $\{S\}$ denotes the set of covariates implicitly defining the subgroups in which the instrument is valid according to Card. $\{R\}$ contains all remaining covariates (for details see main text). Consistent with Card's argument, the null hypothesis of instrument validity cannot be rejected when all covariates are included; see column $\{R, S\}$. Similarly, when only the six covariates $\{S\}$ are included the instrument validity can also not be rejected; see column $\{S\}$. Dropping all variables $\{S\}$ and keeping only those in $\{R\}$, the test rejects instrument validity.*

Acemoglu et al. (2008) investigate the causal relationship between the level of economic development and democracy by using {past savings rates} as an instrumental variable for {economic development} (in a simple capital accumulation growth model, higher savings rates cause more economic growth, see e.g. Mankiw et al., 1992). As acknowledged by the authors, the validity of the instrument is debatable as, e.g., saving rates might be correlated with anticipated regime changes. Nevertheless, the authors claim that it seems "plausible" (p. 822) that saving rates do not have a direct effect on the culture of democracy.

We use the data provided by Acemoglu et al. (2008) to evaluate the validity of {past savings rates} as an instrument. The data comprises information about the per capita Gross Domestic Product (the cause variable), Freedom House democracy index (the effect variable), aggregate saving rate (the proposed instrument), a number of additional control variables (level of education, population size, median population age, labor share, country and year dummies) for 85 countries. We cannot reject the validity of {past savings rates} as an instrument for {economic development} since the p-value for the hypothesis H_0 : IV is valid is 0.520. See Table 4 in Appendix J. Thus, we can substantiate the narrative justification for the instrument given by Acemoglu et al. (2008).

6.3 Interpretation of PIM in the preceding case studies

Throughout, we have stressed that PIM is the crucial underlying idea of the proposed IV validity test. Therefore, it is instructive to discuss what PIM amounts to in the two preceding empirical applications.

First, how can PIM be interpreted in the returns to education study (Section 6.1)? Spelling it out, PIM states that the mechanism that translates general work ethic or parental socio-economic background (and other variables that are captured by U) into educational attainment is independent of the mechanism that translates educational attainment into job earnings. Consider a scenario where a hypothetical policymaker were to make college admissions entirely independent of students' parental socio-economic status, i.e. the mechanism translating U into X would change. If this intervention in a specific societal mechanism were

to alter the causal effect of educational attainment on job earnings (e.g. because it changed how much employers value analytical thinking skills acquired in college), PIM would be violated. However, this seems unlikely at least in the short run and, therefore, assuming the independence of the described mechanisms seems reasonable in the case at hand.

Second, how can PIM be interpreted in the study on democratic development (Section 6.2)? PIM, here, is the assumed independence of the following two mechanisms: first, there is the mechanism that translates the level of educational attainment (and other confounding factors captured by U) into economic development as measured by GDP growth. One can think of a higher educational level in an economy causing more economic development because it opens up possibilities in high-growth technology sectors. Second, there is the mechanism that translates economic development into democratic development. What is the nature of this latter mechanism? Oversimplifying, over the course of economic development many institutions, e.g. a better organization of the middle class through unionization, evolve that cause democratic development (Lipset, 1959).[9] Consider a hypothetical social engineer that would alter the direct causal effect of educational level in an economy on economic output by, e.g., instituting broadly accessible labor markets. PIM amounts to assuming that such a change would not alter the mechanistic relation between economic development and democratization. Assessing whether those two mechanisms can indeed be thought of as independent would require a more extensive discussion about what kind of processes are at play that translate educational level into economic growth and economic growth into democratic development, respectively. This is beyond the scope of this paper.

7. Conclusion

Since the justification of IV assumptions is in practice seldom statistically-grounded and often relies on controversial context-specific arguments, it is pertinent to provide methods to evaluate IV validity empirically. The proposed method leverages statistical traces of confounding in observed data, which can be measured with the method laid out in Janzing and Schölkopf (2018a), to test whether a potential instrument is valid. It provides a novel way to test IV validity, which unlike previous work does not rely on the testable implications derived by Balke and Pearl (1997) nor on higher-order moment conditions. Thus, it constitutes a novel approach to evaluating IV validity and adds to the literature that employs the Principle of Independent Mechanisms to address practical causal inference problems.

Using the method by Janzing and Schölkopf (2018a), a degree of confounding is estimated for the model where the treatment variable of interest is instrumented with the possibly invalid instrument. The estimate of the degree of confounding thus obtained is model-wide and not informative of confounding of a single covariate, i.e. the instrumented treatment variable in this case. The first main contribution of this paper is to show how to construct a synthetic, unconfounded variable to estimate a counterfactual degree of confounding that would be obtained if the instrument were valid. The second main contribution of this paper is to show that comparing the counterfactual degree of confounding to the observed degree of confounding is informative about instrument validity.

9. Within the context of this paper, we cannot do justice to the large body of academic work on the theory of modernization that deals with the relation between democratic political structures and economic development. Instead, we merely cite the seminal argument, which has spawned much of that literature.

Monte Carlo studies show that the proposed method has high accuracy. Its AUC levels reach from around 0.7 when the number of observations, covariates, and degree of violation of crucial IV assumptions is low to levels close to 1 when the number of observations, covariates and degree of violation of validity assumptions increases. Despite different theoretical approaches, we compare the performance of our test to the one proposed by Kitagawa (2015). Our test performs favorably in the linear setting studied here (note though that Kitagawa's approach is also applicable in nonparametric models). We document the feasibility of the proposed test in two empirical applications. First, we show that the test can corroborate an argument for the validity of college proximity as an instrument for educational attainment due to Card (1995). Second, the validity of past saving rates as an instrument for economic development cannot be rejected (Acemoglu et al., 2008). Moreover, we show the robustness of the procedure to two different ways of estimating the degree of confounding and two different ways of creating the synthetic variable.

Acknowledgments

This paper is developed from the chapter "Structural Autonomy and Instrument Validity" of my doctoral dissertation, which I defended at the Free University of Berlin in July 2020. I am grateful for helpful comments and suggestions by Frederick Eberhardt, Uri Shalit, Christoph Breunig, Carsten Schröder, Johannes König, Max Schäfer, and Dominik Janzing. Remaining errors are mine.

Appendix A. Proof: Relation between $(\kappa_i - \kappa_s)$ and IV validity

First, we repeat the definition of a valid IV.

Definition 1. *A variable Z is called a valid instrumental variable if and only if it fulfills Assumptions 1-3.*

For convenience, we reproduce the reduced form model that forms the starting point for the test in Section 3:

$$Y = \{\mathbf{X}, T_{\text{IV}}\} \begin{pmatrix} \beta \\ \tau \end{pmatrix} + c\tilde{U} + \varepsilon \tag{25}$$

$$\{\mathbf{X}, T_{\text{IV}}\} = \mathbf{E} + \tilde{U} \begin{pmatrix} \mathbf{b} & b_{T_{\text{IV}}} \end{pmatrix}. \tag{26}$$

Each element of the vector $\begin{pmatrix} \mathbf{b} & b_{T_{\text{IV}}} \end{pmatrix} = \begin{pmatrix} b_1 & \dots & b_d & b_{T_{\text{IV}}} \end{pmatrix}$ parameterizes the confounding of the corresponding dimension of $\{\mathbf{X}, T_{\text{IV}}\}$, e.g. $X_1 = E_1 + \tilde{U}b_1$. If Z is a valid IV, the instrumented treatment variable T_{IV} is unconfounded, and $b_{T_{\text{IV}}} = 0$.

For convenience, we reproduce the definition of κ_i and κ_s here and introduce some placeholders.

$$\kappa_s = \frac{\overbrace{\left\| c_s \Sigma^{-1}_{\mathbf{X}T_s} \begin{pmatrix} \mathbf{b} \\ b_{T_s} \end{pmatrix} \right\|^2}^{\bar{c}_s}}{\underbrace{\left\| \begin{pmatrix} \beta \\ \tau_s \end{pmatrix} \right\|^2}_{\bar{a}_s} + \left\| c_s \Sigma^{-1}_{\mathbf{X}T_s} \begin{pmatrix} \mathbf{b} \\ b_{T_s} \end{pmatrix} \right\|^2} \tag{27}$$

$$\kappa_i = \frac{\overbrace{\left\| c \Sigma^{-1}_{\mathbf{X}T_{\text{IV}}} \begin{pmatrix} \mathbf{b} \\ b_{T_{\text{IV}}} \end{pmatrix} \right\|^2}^{\bar{c}_\tau}}{\underbrace{\left\| \begin{pmatrix} \beta \\ \tau \end{pmatrix} \right\|^2}_{\bar{a}_\tau} + \left\| c \Sigma^{-1}_{\mathbf{X}T_{\text{IV}}} \begin{pmatrix} \mathbf{b} \\ b_{T_{\text{IV}}} \end{pmatrix} \right\|^2} \tag{28}$$

Note that $\tau_s = b_{T_s} = 0$ since we draw T_s independently of Y. By virtue of how T_s is generated, $\Sigma_{\mathbf{X}T_s} = \Sigma_{\mathbf{X}T_{\text{IV}}}$. Under instrument validity, replacing T_{IV} with T_s the relation between Y and $\tilde{U}$ does not change and, therefore, $c_s = c$.

For convenience, we reproduce the Theorem 1 here before proving it.

Theorem 1. *If the instrumental variable is valid, δ is not positive:*

$$\textit{IV valid} \Rightarrow \delta := \kappa_i - \kappa_s \leq 0. \tag{29}$$

Proof If the instrumental variable is valid, $b_{T_{\text{IV}}} = 0$. Then,

$$\begin{aligned}\kappa_i - \kappa_s &= \frac{\left\| c\Sigma^{-1}_{\mathbf{X}T_{\text{IV}}} \begin{pmatrix} \mathbf{b} \\ b_{T_{\text{IV}}} \end{pmatrix} \right\|^2}{\left\| \begin{pmatrix} \beta \\ \tau \end{pmatrix} \right\|^2 + \left\| c\Sigma^{-1}_{\mathbf{X}T_{\text{IV}}} \begin{pmatrix} \mathbf{b} \\ b_{T_{\text{IV}}} \end{pmatrix} \right\|^2} - \frac{\left\| c_s\Sigma^{-1}_{\mathbf{X}T_s} \begin{pmatrix} \mathbf{b} \\ b_{T_s} \end{pmatrix} \right\|^2}{\left\| \begin{pmatrix} \beta \\ \tau_s \end{pmatrix} \right\|^2 + \left\| c_s\Sigma^{-1}_{\mathbf{X}T_s} \begin{pmatrix} \mathbf{b} \\ b_{T_s} \end{pmatrix} \right\|^2} \\ &= \frac{\bar{c}}{\left\| \begin{pmatrix} \beta \\ \tau \end{pmatrix} \right\|^2 + \bar{c}} - \frac{\bar{c}}{\left\| \begin{pmatrix} \beta \\ \tau_s \end{pmatrix} \right\|^2 + \bar{c}} \leq 0 \end{aligned} \tag{30}$$

where $\bar{c} = \left\| c\Sigma^{-1}_{\mathbf{X}T_{\text{IV}}} \begin{pmatrix} \mathbf{b} \\ b_{T_{\text{IV}}} \end{pmatrix} \right\|^2 = \left\| c_s\Sigma^{-1}_{\mathbf{X}T_s} \begin{pmatrix} \mathbf{b} \\ b_{T_s} \end{pmatrix} \right\|^2$ because $\Sigma_{\mathbf{X}T_{\text{IV}}} = \Sigma_{\mathbf{X}T_s}$ and $b_{T_{\text{IV}}} = b_{T_s}$. In other words, neither $\bar{c}_\tau$ nor $\bar{c}_s$ contain τ or τ_s, which are the only quantities that differ between κ_s and κ_i if the IV is valid. The last inequality is due to the fact that $\tau_s = 0$ by construction. Therefore, it follows

$$\text{IV valid} \Rightarrow \delta \leq 0.$$

■

By contrapositive, this result implies that if $\delta > 0$, the instrumental variable is invalid. Thus, the proposed test evaluates the null hypothesis H_0 : IV valid.

Appendix B. Detailed explanation of Algorithm 2

Since the creation of the synthetic treatment variable in Algorithm 2 is an important part of the proposed instrument validity test, we provide a detailed explanation of each step of the algorithm here (the numbering in what follows corresponds to the line numbers in Algorithm 2).[10] For convenience, we iterate that we want to generate a synthetic variable T_s such that

$$\text{Cov}(X_i, T_s) = \text{Cov}(X_i, T_{\text{IV}}) \; \forall i \in \{1, \ldots, d\}. \tag{31}$$

1. The elements of the vector ρ contain the covariance that we want the resulting synthetic variable T_s to have with the respective dimension of the observed data $\mathbf{X}$. So, $\rho_i = \text{Cov}(X_i, T_{\text{IV}})$.

2. In this step, we generate a random vector W unrelated to $\mathbf{X}$, U and Y (and therefore surely unconfounded) from which we will regress out all variation it has with $\mathbf{X}$ by chance (Step 3) to then add parts of $\mathbf{X}$ in a specific way that ensures we construct a random variable with the desired covariance structure.

3. We take out all variation in W that can be explained by $\mathbf{X}$ by regressing W on $\mathbf{X}$ and computing the resulting residuals: η. The vector η is orthogonal to all columns of $\mathbf{X}$ by the mechanics of OLS.

10. Algorithm 2 builds on an idea by CrossValidated user `whuber`, see `https://tinyurl.com/syntheticT`.

We construct T_s as a linear combination of η and $\mathbf{X}$,

$$T_s = s\eta + \sum_{i=1}^{d} \alpha_i X_i, \tag{32}$$

where α_i are scalar weights. We want to find α_i such that (31) holds, i.e.,

$$\begin{aligned} \operatorname{Cov}(X_i, T_s) &= \operatorname{Cov}\left(X_i, s\eta + \sum_{i=1}^{d} \alpha_i X_i\right) && (33)\\ &= \operatorname{Cov}(X_i, s\eta) + \operatorname{Cov}\left(X_i, \sum_{i=1}^{d} \alpha_i X_i\right) && (34)\\ &= \operatorname{Cov}\left(X_i, \sum_{i=1}^{d} \alpha_i X_i\right) = \rho_i && (35) \end{aligned}$$

It turns out that rewriting the linear combination in terms of the dual of $\mathbf{X}$, $\mathbf{X}_{\text{dual}}$, i.e.,

$$\sum_{i=1}^{d} \alpha_i X_i = \sum_{j=1}^{d} \gamma_j X_{\text{dual},j} \tag{36}$$

is more convenient due to the following relation:

$$\operatorname{Cov}\left(X_i, \sum_{j=1}^{d} \gamma_j X_{\text{dual},j}\right) = \sum_{j=1}^{d} \gamma_j \operatorname{Cov}(X_i, X_{\text{dual},j}) = \gamma_i \tag{37}$$

which holds because $\operatorname{Cov}(X_i, X_{\text{dual},j}) = 1$ for $i = j$ and $\operatorname{Cov}(X_i, X_{\text{dual},j}) = 0$ for $i \neq j$.

In words, the covariance of the linear combination in terms of the dual of $\mathbf{X}$ (right-hand side of eq. (36)) and X_i is equal to the respective weight of $X_{\text{dual},i}$ in that linear combination. This is convenient because it allows us to construct T_s from the linear combination of the dual versions of X_i weighted by the target covariance ρ_i, respectively.

4. Compute the Singular Value Decomposition, which serves as input for the following step.

5. Using the results from the Singular Value Decomposition, compute $\mathbf{X}_{\text{dual}}$ as described in Step 5 of the algorithm.

6. We want to ensure that $\operatorname{Var}(T_s) = 1$. We can do that by choosing s in eq. (32) appropriately. To solve for s, first observe that

$$\begin{aligned} 1 &= \operatorname{Var}(T_s) && (38)\\ &= \operatorname{Var}(s\eta + \sum_{j=1}^{d} \gamma_j X_{\text{dual},j}) && (39)\\ &= s^2 \operatorname{Var}(\eta) + \operatorname{Var}(\mathbf{X}_{\text{dual}}\rho) && (40)\\ &= s^2 \operatorname{Var}(\eta) + \rho^\top \operatorname{Cov}(\mathbf{X}_{\text{dual}})\rho. && (41) \end{aligned}$$

This implies that

$$s = \left(\frac{1 - \rho^\top \text{Cov}(\mathbf{X}_{\text{dual}})\rho}{\text{Var}(\eta)} \right)^{1/2}. \tag{42}$$

7. Having established that we can use ρ_i as weights in the construction of T_s as a linear combination if we use the dual of $\mathbf{X}$, we can write

$$T_s = \mathbf{X}_{\text{dual}} \times \rho + s \times \eta \tag{43}$$

Note that we not require T_s to follow a conditionally Gaussian model. All subsequent steps in Algorithm 1 merely require the covariance structure of T_s to fulfill eq. (31), regardless of the distribution of T_s.

Appendix C. Further results on the empirical application

The results in Section 6 suggest a question about what happens when different subsets of $\{S\}$ are included as covariates. Therefore, we run the proposed test for all subsets of covariates $\{S\}$. For rows showing a pseudo-p-value > 0.05, the average number of covariates from $\{S\}$ that is included is 3.6. For rows showing a pseudo-p-value ≤ 0.05, the average number of covariates from $\{S\}$ that is included is 1.9. The number of variables included from $\{S\}$ correlates positively with the pseudo-p-value, which is consistent with the main message of Section 6, namely that the inclusion of $\{S\}$ renders the instrument valid. The pseudo-p-value reported in the first column of Table 1 appears in the row in which all covariates from $\{S\}$ are included. This row is not associated with the largest pseudo-p-value. However, given the empirical distribution of the pseudo-p-value under H_0 discussed in Section 5, this is not to be expected.

Appendix D. Robustness to rescaling

As mentioned, a drawback of the JS methodology to estimate a degree of confounding is that it is theoretically not robust to transformations of the data as this introduces a dependence of the parameter vector and the covariance matrix of the covariates. However, the proposed method relies on a comparison of *two* κs. Since both would be affected by transformations in the same way, their difference (which the test relies on) is not affected by transformations. To corroborate this argument, we apply typical data transformation to the data generated in the Monte Carlo study (Regime 1: violation of exclusion restriction) and compare pseudo-p-values for both transformed and untransformed data.

Logarithmic transformations are frequently used in many domains, e.g. in economics where income levels are usually rescaled with logarithms. To avoid infinitely large values after transforming the simulated data, we need to add an additional step to the data generating process in Algorithm 3, which ensures that all values lie above 1. In particular, we transform $\mathbf{X}$, as defined in (47), by

$$\mathbf{X} := \mathbf{X} - \min(\min(\mathbf{X}), 0) + \mathbb{1}(\min(\mathbf{X}) < 0) \tag{44}$$

and we replace Y as defined in eq. (48) by

$$Y := Y - \min(\min(Y), 0) + \mathbb{1}(\min(Y) < 0) \tag{45}$$

pseudo-p	ethnicity dummy	father's educational level	living in South 1966	living in South 1976	urban residence 1966	urban residence 1976	no of covariates	pseudo-p	ethnicity dummy	father's educational level	living in South 1966	living in South 1976	urban residence 1966	urban residence 1976	no. of covariates
0.288	0	0	0	1	1	1	3	0.122	0	1	1	1	0	0	3
0.258	0	1	1	1	1	1	5	0.122	1	0	1	0	1	1	4
0.254	0	1	0	1	1	1	4	0.096	1	0	1	1	0	1	4
0.250	0	0	1	1	1	1	4	0.094	1	0	1	1	0	0	3
0.246	1	0	0	1	1	1	4	0.090	1	1	1	1	0	1	5
0.244	0	0	1	1	1	0	3	0.088	1	1	1	1	0	0	4
0.242	0	1	1	1	1	0	4	0.084	1	0	0	1	1	0	3
0.222	0	0	1	0	1	0	2	0.080	0	0	0	1	1	0	2
0.212	1	1	0	1	1	1	5	0.076	0	1	0	1	1	0	3
0.212	1	0	1	1	1	1	5	0.058	1	0	0	0	1	0	2
0.204	0	0	1	0	1	1	3	0.058	1	1	0	1	1	0	4
0.184	0	1	1	0	1	0	3	0.042	0	0	0	0	1	0	1
0.180	0	1	0	1	0	1	3	0.040	0	1	0	1	0	0	2
0.180	1	1	1	1	1	1	6	0.040	0	1	0	0	1	0	2
0.176	0	1	1	0	1	1	4	0.040	1	1	0	0	1	0	3
0.174	0	0	1	1	0	1	3	0.034	0	0	0	1	0	0	1
0.172	0	0	0	0	1	1	2	0.024	1	1	0	1	0	0	3
0.170	0	1	0	0	1	1	3	0.022	1	0	0	1	0	0	2
0.170	1	1	1	1	1	0	5	0.006	1	0	0	0	0	1	2
0.160	0	0	0	1	0	1	2	0.004	0	0	1	0	0	1	2
0.156	0	1	1	1	0	1	4	0.002	0	1	0	0	0	0	1
0.150	1	0	1	1	1	0	4	0.002	0	0	1	0	0	0	1
0.148	1	1	1	0	1	0	4	0.002	0	1	1	0	0	0	2
0.148	1	1	0	1	0	1	4	0.000	1	0	0	0	0	0	1
0.148	1	1	0	0	1	1	4	0.000	0	0	0	0	0	1	1
0.144	1	0	1	0	1	0	3	0.000	1	1	0	0	0	0	2
0.140	0	0	1	1	0	0	2	0.000	1	0	1	0	0	0	2
0.140	1	0	0	1	0	1	3	0.000	0	1	0	0	0	1	2
0.130	1	1	1	0	1	1	5	0.000	1	1	1	0	0	0	3
0.128	1	0	0	0	1	1	3	0.000	1	1	0	0	0	1	3

Table 2: *Each row of this Table shows the pseudo-p-value for the proposed test when the subset of $\{S\}$ indicated in the respective columns is included as covariates (a '1' denotes inclusion). $\{R\}$ is included for all rows. Column 'no. of covariates' shows how many covariates from $\{S\}$ are included. The results are ordered by pseudo-p-value in descending order. Note that the bottom three rows of the Table are omitted to save space.*

where $\mathbb{1}$ is the indicator function. These transformation ensure that all values lie above 1 and logarithmic transformations do not lead to infinitely large values.

We then implement the following three data transformations:

1. $X_1 := \log(X_1)$, and $X_2 := X_2^2$
2. $Y := \log(Y)$, $X_1 := \log(X_1)$, and $X_2 := X_2^2$
3. $Y := \log(Y)$, $X_1 := \log(X_1)$, and $X_2 := \log(X_2)$

For the original data and for each of the three transformations, we implement the algorithm described in the main text and compare the pseudo-p-values that result. In Figure 10, we show scatter plots of pseudo-p-values for each data transformation against those of the original data. For each data transformation, the pseudo-p-values correlate strongly with those from the original data.

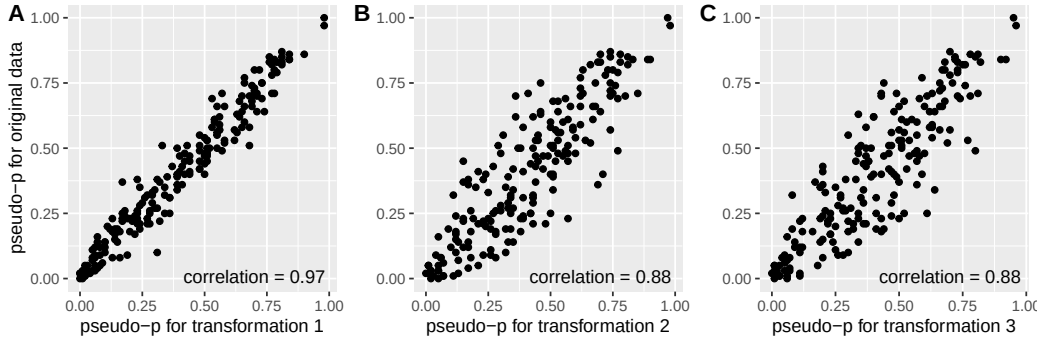

Figure 10: *Scatter plot of pseudo-*p*-values. This Figure shows scatter plots of pseudo-p-values estimated based on transformed data against pseudo-p-values estimated based on the original data. Each panel corresponds to one transformation of the data. The p-values remain largely invariant with each scatter plot displaying a correlation of about 0.9. This is evidence for the robustness of the proposed test for instrument validity with respect to rescaling of the data.* $n = 1000$*,* $d = 20$*,* $\omega_1 = 0.3$*,* $\omega_2 = 0.3$*,* $Var(\varepsilon_Y) = Var(\varepsilon_T) = 1$*.*

Appendix E. Further results for Simulation Regime 1

In this section we provide further simulation results for Simulation Regime 1 (Violation of the exclusion restriction) to show robustness of the results for different variances of the error distributions, $\sigma_Y^2 \in \{0.5, 1, 1.5\}$. Figure 11 is similar to Figure 5 in the main text but shows how the results change for various levels of σ_Y^2. The performance of the test deteriorates slightly when the variance increases.

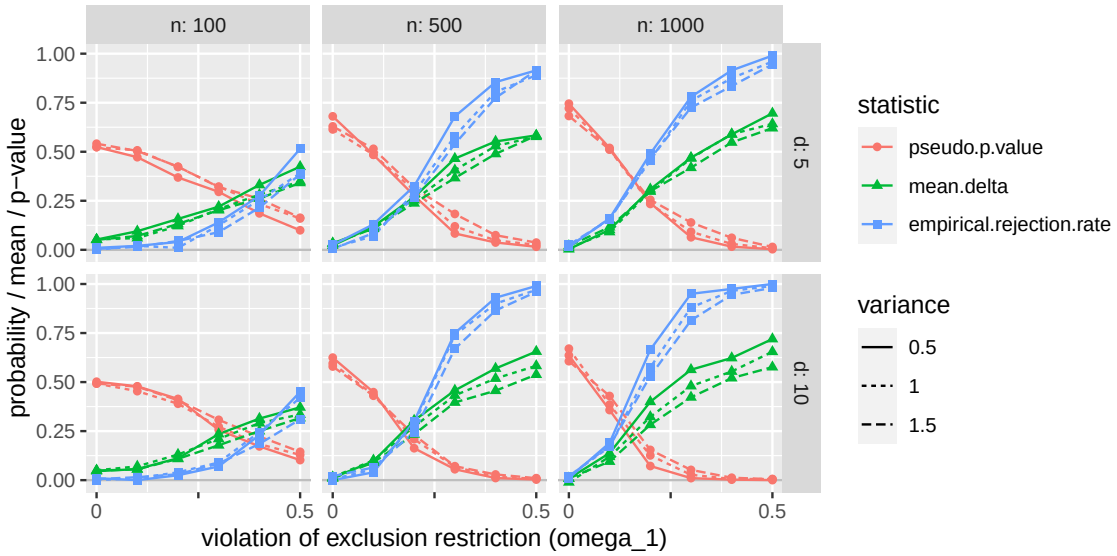

Figure 11: *Simulation results: pseudo-p-values, δ_B, and empirical rejection rate as a function of ω_1 for various levels of σ_Y^2 (denoted 'variance' in the legend). This Figure shows averages over all M Monte Carlo draws of the p-value, δ_B, and the empirical rejection probability (based on the p-value with threshold parameter $\alpha = 0.05$) as a function of the degree of violation of the exclusion restriction (ω_1), by number of covariates d and number of observations n. $\omega_2 = 0.3$. The results described in the main text deteriorate only slightly when σ_Y^2 increases.*

Appendix F. Simulation Regime 2: Violation of Exchangeability Assumption

For the simulations to test whether the algorithm can detect confounding of the instrument stemming from a violation of the exchangeability assumption, we generate data according to Algorithm 4. Figures 12 and 13 show results for the simulations for the violation of the exchangeability assumption. The test performs well also for this violation of IV validity. Note that the degree of endogeneity of the instrument is not directly comparable to Simulation Regime 1 since ω_1 enters the simulation inside an indicator function for Simulation Regime 2.

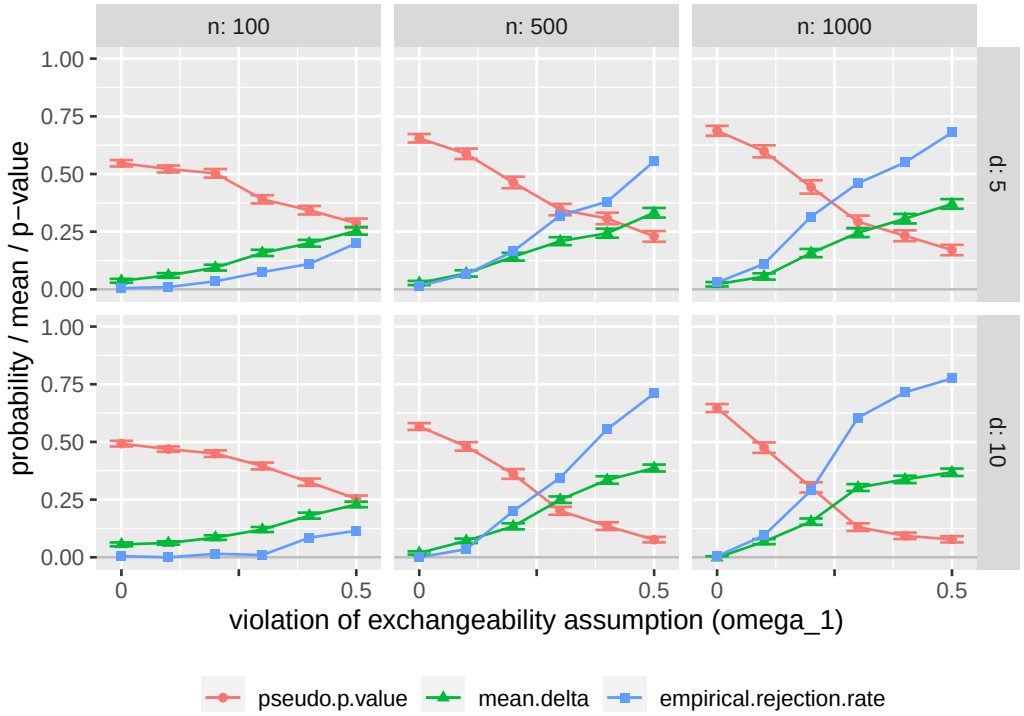

Figure 12: *Simulation results: pseudo-p-values, δ_B, and empirical rejection rate as a function of ω_1. This Figure shows the pseudo-p-value, δ_B, and the empirical rejection probability (based on the pseudo-p-value with threshold parameter $\alpha = 0.05$) as a function of violation of the exchangeability assumption, by number of covariates, (d), and number of observations (n). δ_B rises with the degree of confounding, as does the pseudo-p-value. Consequently, the empirical rejection probabilities go down to zero indicating that, if the degree of condounding is sufficiently high, the test does not reject the null of endogeneity.*

Input: number of observations n, number of covariates d, variance of the structural errors σ^2; two parameters specifying relation between instrument, treatment and unobserved confounder: ω_1: endogeneity of Z, ω_2: the relevance of the instrument Z

Output: a simulated data set of n observations of outcome variable, covariates, treatment variable, instrument $\mathcal{D} = \{Y, \mathbf{X}, T, Z\}_{i=1}^{n}$

1 **Generation of structural errors**: ε_Y and ε_T, are drawn from

$$\begin{pmatrix} \varepsilon_Y \\ \varepsilon_T \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right) \tag{46}$$

2 **Generation of instrument** Z: Let $\varepsilon_Z \sim \mathcal{N}(0,1)$ and $Z = \mathbb{1}(\varepsilon_Z + \omega_1 \varepsilon_Y > 0)$ where $\mathbb{1}$ is the indicator function.

3 **Generation of covariates** $\mathbf{X}$: Draw d eigenvalues from a uniform distribution $\lambda_i \sim \mathcal{U}(0.5, 1.5)$ which populate the diagonal of a $(d) \times (d)$ matrix Λ. Then draw a random orthonormal matrix $\mathbf{O}$ of dimension (d), set $\Sigma = \mathbf{O}\Lambda\mathbf{O}^\top$ and draw $\mathbf{X}_{\text{temp}}$ from a multivariate normal distribution $\mathbf{X}_{\text{temp}} \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

4 Draw a random (d)-dimensional vector from a normal distribution: $\beta_{c,temp} \sim \mathcal{N}(0,1)$ and, to keep the variance of Y comparable for various d, normalize $\beta_c = \beta_{c,temp} / \left\| \beta_{c,temp} \right\|$. With these ingredients set

$$\mathbf{X} = \mathbf{X}_{\text{temp}} + \varepsilon_Y \beta_c^\top. \tag{47}$$

5 **Generation of treatment** T: To induce dependence of the treatment on the set of covariates, first draw the d-dimensional vector $\beta_{T,temp}$ populated with draws from a $\mathcal{N}(0,1)$, $\beta_{T,temp} \sim \mathcal{N}(0,1)$ and set $\beta_T = \left(\beta_{T,temp}\right) / \left\| \left(\beta_{T,temp}\right) \right\|$ to keep the relative influence of $\mathbf{X}$ on T comparable for various d.

6 Generate T, as $T = \mathbb{1}\left(\mathbf{X}\beta_T^\top + \omega_2 Z + \varepsilon_T > T' \right)$ where T' is the mean of $\mathbf{X}\beta_T^\top + \varepsilon_T$ and $\mathbb{1}$ is the indicator function.

7 **Generation of the outcome variable** Y: First generate a random d-dimensional vector $\beta_{\text{temp}} \sim \mathcal{N}(0,1)$. To keep the variance of Y comparable for various d, set $\beta = \left(\beta_{\text{temp}}\right) / \left\| \left(\beta_{\text{temp}}\right) \right\|$. The true causal effect of the treatment variable is set to $\tau = 1$. Finally, generate outcome Y as

$$Y = \mathbf{X}\beta^\top + \tau T + \varepsilon_Y. \tag{48}$$

Normalization of data To keep the binary nature of T, we normalize the data to have equal variance and equal mean as T (as opposed to normalizing all variables to have zero mean and variance equal to one).

Algorithm 4: Simulation of Violation of Exchangeability Assumption

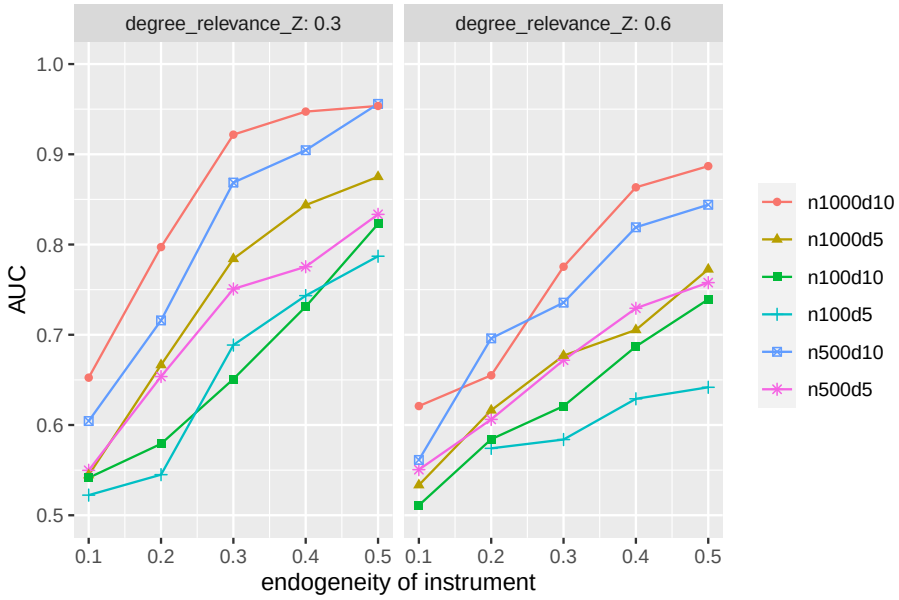

Figure 13: *AUC curves for violations of the exchangeability assumption. This figure shows the area under the ROC curve (AUC) as a function of the degree of violation of the exchangeability assumption, for various combinations of number of covariates, d, and number of observations, n. Underlying test statistic is the pseudo-*p*-value. The test achieves high AUC levels of close to the perfect score of 1 for large n and d.*

Appendix G. Simulation for the illustration of PIM

The illustration in Figure 4 is based on the following simulation.

First, construct a covariance matrix Σ as follows. Draw $d+1$ eigenvalues

$$\lambda \sim \mathcal{U}(0.5, 1.5)$$

which populate the diagonal of a matrix V. Then we draw a random orthogonal matrix L and set $\Sigma = VLV^\top$. We multiply each element in the last row and last column of Σ by 5 to induce more unexplained variation in Y. For the unconfounded case, we replace the last row and last column of Σ with zeroes but leave the $(d+1, d+1)$ entry untouched:

$$\begin{aligned} \text{confounded: } S_c &= \Sigma_{d+1\times d+1} \\ \text{unconfounded: } S_u &= \begin{pmatrix} \Sigma_{(1:d)\times(1:d)} & \mathbf{0} \\ \mathbf{0} & \sigma_{d+1\times d+1} \end{pmatrix} \end{aligned} \tag{49}$$

We simulate data by drawing the structural error term ε_Y and $\mathbf{X}$ from a jointly normal distribution

$$\begin{pmatrix} \mathbf{X} \\ \varepsilon_Y \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, S_i) \tag{50}$$

where $i \in \{c, u\}$.

Next, draw the d-dimensional true parameter vector

$$\beta \sim \mathcal{N}(\mathbf{0}, diag(1))$$

and divide each element of β by $d^{0.5}$ (to keep the variance of Y comparable for different d).

Finally, set

$$Y = \mathbf{X}\beta + \varepsilon_Y. \tag{51}$$

We estimate $\hat{\beta}$ by OLS.

Appendix H. Janzing and Schölkopf (2018a) in a nutshell

Janzing and Schölkopf (2018a) propose a method to estimate the degree to which an observed statistical relationship between a multidimensional set of covariates, $\mathbf{X}$, and an outcome variable Y is due to the causal influence of $\mathbf{X}$ on Y or due to an unobserved confounder influencing both $\mathbf{X}$ and Y. This section does not contain new results.

Section 4 contains an illustration why the orientation of a parameter vector with respect to the eigenspaces of the corresponding $\Sigma_{\mathbf{XX}}$ contains a confounding signal. JS propose a method to measure deviations from the generic orientation to estimate the *degree of confounding* in multivariate linear models. Technically, generic orientation is instantiated as the equivalence of two spectral measures of $\Sigma_{\mathbf{XX}}$: first, the unweighted spectral measure (called tracial spectral measure and denoted $\mu^{Tr}_{\Sigma_{\mathbf{XX}}}$), and second, the spectral measure weighted by a vector such as a parameter vector β (called vector-induced spectral measure denoted $\mu_{\Sigma_{\mathbf{XX}},\beta}$)[11]:

$$\text{generic orientation of } \beta \text{ with respect to } \Sigma_{\mathbf{XX}} \Leftrightarrow \mu^{Tr}_{\Sigma_{\mathbf{XX}}} \simeq \mu_{\Sigma_{\mathbf{XX}},\beta}. \tag{52}$$

Ideally, one would check whether the spectral measure induced by the estimated parameter vector is equivalent to that induced by the true parameter vector. However, the latter is not estimable from observed data. Nevertheless, the equivalence of the tracial spectral measure and that induced by the true parameter vector makes it possible to compare the spectral measure induced by the estimated, and possibly biased, vector to the estimable tracial spectral measure to infer a degree of confounding.

The crucial result in JS is that the computable spectral measure induced by the estimated (and possibly biased) parameter vector, $\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}}$, can be decomposed into one part that is due to confounding and a second part that represents genuine causation. More specifically, $\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}}$ can be decomposed into the spectral measure induced by the true parameter vector and that induced by the bias of the estimated parameter vector from the true parameter vector. The relative sizes of these two components define the degree of confounding κ:

$$\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}} \simeq (1-\kappa)\, \mu_{\Sigma_{\mathbf{XX}},\beta} + \kappa\, \mu_{\Sigma_{\mathbf{XX}},(\hat{\beta}-\beta)}. \tag{53}$$

11. We use $\simeq$ in this and the following expressions in this subsection to indicate that the following statements are not precise in the sense that we do not explicitly state the types of and rates of convergence as well as conditions for convergence. See the following subsections in this Appendix H for details.

κ ranges from 0 (no confounding) to 1 (observed statistical relation is fully due to confounding). Without confounding,

$$\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}} \simeq \mu_{\Sigma_{\mathbf{XX}},\beta} \simeq \mu^{Tr}_{\Sigma_{\mathbf{XX}}}, \tag{54}$$

i.e. $\hat{\beta}$ is generically oriented.

Still, $\mu_{\Sigma_{\mathbf{XX}},\beta}$ and $\mu_{\Sigma_{\mathbf{XX}},(\hat{\beta}-\beta)}$ cannot be determined since they involve the unknown true β. However, the estimable $\mu_{\Sigma_{\mathbf{XX}},\hat{\beta}}$ can be parameterized by a two-parametric family of probability measures. The algorithm proposed by JS finds those two parameter values that minimize the distance between the two-parametric estimate and the observed spectral measure induced by the estimated (and possibly) biased parameter vector. One of the parameters is κ.

The next subsection contains the formal steps needed to achieve this result.

H.1 The set-up

Consider the following linear structural equation model:

$$\mathbf{X} = \mathbf{b}U + \mathbf{E} \tag{55}$$

$$Y = \mathbf{X}^\top \mathbf{a} + cU^\top + \varepsilon \tag{56}$$

where Y is the $n \times 1$ outcome vector, $\mathbf{a}$ is the $d \times 1$ causal parameter vector of interest. $\mathbf{X}$ is a $d \times n$ matrix of covariates. The confounder U is a $1 \times n$ vector. $\mathbf{b}$ is a $d \times 1$ parameter vector. $\mathbf{E}$ is a $d \times n$ matrix of zero-mean errors drawn independently from u. ε is a $n \times 1$ vector of errors. c is a scalar. Without loss of generality, u is assumed to have unit variance.

After projecting with least-squares in the population, the parameter vector is given by

$$\hat{\mathbf{a}} := \Sigma^{-1}_{\mathbf{XX}} \Sigma_{\mathbf{X}Y}, \tag{57}$$

where Σ denotes covariance matrices. Generally, we are interested in the structural parameter vector $\mathbf{a}$ which represents genuine causal influence. To illustrate, the relation between $\mathbf{a}$ and $\hat{\mathbf{a}}$ consider

$$\begin{aligned}
\Sigma_{\mathbf{X}Y} &= \mathrm{Cov}(\mathbf{X}, Y) = \mathrm{Cov}(\mathbf{b}U + \mathbf{E}, \mathbf{X}^\top \mathbf{a} + cU^\top + \varepsilon) \\
&= (\Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top)\mathbf{a} + c\mathbf{b} \\
\Sigma_{\mathbf{XX}} &= \mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \mathrm{Cov}(\mathbf{b}U + \mathbf{E}, \mathbf{b}U + \mathbf{E}) \\
&= \Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top,
\end{aligned}$$

and therefore

$$\hat{\mathbf{a}} = \mathbf{a} + (\Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top)^{-1} c\mathbf{b} = \mathbf{a} + c\Sigma^{-1}_{\mathbf{XX}} \mathbf{b}. \tag{58}$$

H.2 Genericity assumptions

The idea underlying this method is the Independence between Cause and Mechanism (ICM) postulate (Peters et al., 2017), which states that the causal mechanism, represented by the conditional distribution of effect, Y, given cause, $\mathbf{X}$, $f(Y|\mathbf{X})$, is independent of the marginal distribution of the cause, $f(\mathbf{X})$.

To understand what the ICM amounts to in the case at hand, note that the crucial determinant for $f(\mathbf{X})$ is $\Sigma_{\mathbf{XX}}$, likewise the crucial determinant for $f(Y|\mathbf{X})$ is $\mathbf{a}$. Therefore, Janzing and Schölkopf (2018a) postulate that $\mathbf{a}$ lies in 'generic orientation' relative to $\Sigma_{\mathbf{XX}}$. For instance, since $\mathbf{a}$ is chosen independently of X, and, thus, also the covariance matrix $\Sigma_{\mathbf{XX}}$, $\mathbf{a}$ is not likely to be aligned with its first principal component.[12] We next discuss what the concept of 'generic orientation' amounts to.

In order to make the notion of 'generic orientation' precise, some definitions and results are needed. First of all, assuming that all eigenvalues of a matrix are different from each other (i.e. the matrix is non-degenerate), each such symmetric $d \times d$ matrix A has a unique decomposition

$$A = \sum_{j=1}^{d} \lambda_j \phi_j \phi_j^\top \tag{59}$$

where λ_j denotes the eigenvalues and ϕ_j the corresponding normalized eigenvectors.

The renormalized trace is defined to be

$$\tau(A) := \frac{1}{d}\mathrm{tr}(A) \tag{60}$$

(note that the τ in this notation is unrelated to the treatment effect that it denotes in the main body of the paper).

Definition 2 (tracial spectral measure)**.** *Let A be a real symmetric matrix with non-degenerate spectrum. The tracial spectral measure of A is defined as the uniform distribution over its eigenvalues $\lambda_1, \ldots, \lambda_d$:*

$$\mu_A^{\mathrm{Tr}} := \frac{1}{d} \sum_{j=1}^{d} \delta_{\lambda_j} \tag{61}$$

where δ_{λ_j} denotes the point measure on λ_j.

The tracial measure is a property of a matrix. The vector-induced spectral measure complements the tracial measure by accounting for its relation to an arbitrary d-dimensional vector.

Definition 3 (vector-induced spectral measure)**.** *Given a symmetric $d \times d$ matrix A with associated eigenvalues λ_j and corresponding eigenvectors ϕ_j, the spectral measure induced by an arbitrary vector $v \in \mathbb{R}^d$ is given by*

$$\mu_{A,v} = \sum_{j=1}^{d} \left(v^\top \phi_j\right)^2 \delta_{\lambda_j} \tag{62}$$

where δ_{λ_j} denotes the point measure on λ_j.

Intuitively, $\mu_{A,v}$ describes the squared length of components of a vector projected onto the eigenspace of $\Sigma_{\mathbf{XX}}$. Note that the vector-induced spectral measure of a matrix can

12. To be precise, for the structural model in eqs. (55)-(55), the argument involves a generic orientation of $\mathbf{a}$ and *the eigenspaces* of $\Sigma_{\mathbf{XX}}$.

be represented by two vectors: one which represents the support of the spectral measure, i.e. a list of the eigenvalues in decreasing magnitude and a second composed of weights corresponding to the eigenvalues. For tracial spectral measures the weight vector is $w = (1/d, \ldots, 1/d)$ representing the uniform weight of the eigenvalues.

The following result for the relation between tracial and vector-induced spectral measure can be shown.

Lemma 4. *For a sequence of positive semi-definite $d \times d$ matrices $(A_d)_{d\in\mathbb{N}}$ with finite norm whose spectral measure converges weakly to some probability measure μ^∞, i.e.*

$$\mu_A^{\mathrm{Tr}} \to \mu^\infty,$$

and a sequence of d-dimensional vectors $(\mathbf{v}_d)_{d\in\mathbb{N}}$ drawn randomly from a sphere of radius r,

$$\mu_{A_d,\mathbf{v}_d} \to r^2\mu^\infty. \tag{63}$$

To formally justify the proposed method to estimate a degree of confounding, JS consider a sequence of generating models for an increasing dimensionality d whose properties are summarized in the following Assumption.

Assumption 5 (Rotation-invariant priors)**.** *The structural parameters of the structural model in eqs.* (55) *and* (56) *are generated as follows:*

1. *The covariance matrix of $\mathbf{E}$ is a uniformly bounded sequence of positive semi-definite $d \times d$-matrices, $(\Sigma^d_{\mathbf{EE}})_{d\in\mathbb{N}}$, such that their tracial measure converges weakly to some probability measure μ^∞, which describes the asymptotic distribution of eigenvalues.*

2. *The vector $(\mathbf{a}_d)_{d\in\mathbb{N}} \in \mathbb{R}^d$ is drawn uniformly at random from a sphere with a fixed radius $r_{\mathbf{a}}$.*

3. *The vector $(\mathbf{b}_d)d \in \mathbb{N} \in \mathbb{R}^d$ is drawn independently from $\mathbf{a}$ and uniformly at random from a sphere with a fixed radius $r_{\mathbf{b}}$. The scalar c is fixed for all d.*

Then $\Sigma^d_{\mathbf{XX}} = \Sigma^d_{\mathbf{EE}} + \mathbf{b}_d\mathbf{b}_d^T$ and $\hat{\mathbf{a}}_d = \mathbf{a}_d + c\left(\Sigma^d_{\mathbf{XX}}\right)^{-1}\mathbf{b}_d$, for which the following Theorem is proven

Theorem 6 (Asymptotic spectral measures)**.** *Given the rotation-invariant priors in Assumption 5, it holds that*

$$\mu_{\Sigma^d_{\mathbf{XX}},\mathbf{a}_d} \to r^2_{\mathbf{a}}\mu^\infty \text{ (weakly in probability)} \tag{64}$$

$$\mu_{\Sigma^d_{\mathbf{EE}},\mathbf{b}_d} \to r^2_{\mathbf{b}}\mu^\infty \text{ (weakly in probability)} \tag{65}$$

$$\mu_{\Sigma^d_{\mathbf{XX}},\mathbf{a}_d+c(\Sigma^d_{\mathbf{XX}})^{-1}\mathbf{b}_d} - \left(\mu_{\Sigma^d_{\mathbf{XX}},\mathbf{a}_d} + \mu_{\Sigma^d_{\mathbf{XX}},\left((\Sigma^d_{\mathbf{XX}})^{-1}\mathbf{b}_d\right)}\right) \to 0 \text{ (weakly in probability)} \tag{66}$$

.

Given these definitions of spectral and vector-induced spectral measures, Lemma 4, Assumption 5, and Theorem 6, the precise meaning of 'generic orientation' is formalized in the following postulate.

Postulate 7 (Generic orientation of vectors)**.** *Given the structural model in eqs.* (55)-(56) *and a large d, we define 'generic orientation' as:*

1. *Vector* $\mathbf{a}$ *has generic orientation relative to* $\Sigma_{\mathbf{XX}}$ *in the sense that*

$$\mu_{\Sigma_{\mathbf{XX}},\mathbf{a}} \approx \mu^{Tr}_{\Sigma_{\mathbf{XX}}} ||\mathbf{a}||^2 \tag{67}$$

2. *Vector* $\mathbf{b}$ *has generic orientation relative to* $\Sigma_{\mathbf{EE}}$ *in the sense that*

$$\mu_{\Sigma_{\mathbf{EE}},\mathbf{b}} \approx \mu^{Tr}_{\Sigma_{\mathbf{EE}}} ||\mathbf{b}||^2. \tag{68}$$

3. *Vector* $\mathbf{a}$ *is generic relative to* $\mathbf{b}$ *and* $\Sigma_{\mathbf{EE}}$ *in the sense that*

$$\mu_{\Sigma_{\mathbf{XX}},\mathbf{a}+c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}} \approx \mu_{\Sigma_{\mathbf{XX}},\mathbf{a}} + \mu_{\Sigma_{\mathbf{XX}},c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}}. \tag{69}$$

The $\approx$-sign is used here because Theorem 6 shows that the vector-induced spectral measures only converge weakly in probability. This postulate is justified by Theorem 6, i.e. the asymptotic behavior of vector-induced spectral measures.

Intuitively, (67) states that 'decomposing $\mathbf{a}$ into eigenvectors of $\Sigma_{\mathbf{XX}}$ yields weights that are close to being uniformly spread over the spectrum.' Equation (68) captures a similar statement for $\mathbf{b}$ and $\Sigma_{\mathbf{EE}}$: the weights of $\mathbf{b}$ are uniformly distributed across the spectrum of $\Sigma_{\mathbf{EE}}$.

Eq. (69) contains a crucial ingredient for the ability to detect confounding: the $\hat{\mathbf{a}}$-induced spectral measure (left-hand-side of (69), recall $\hat{\mathbf{a}} = \mathbf{a} + c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}$) can be decomposed into one part due to the causal vector $\mathbf{a}$ (first summand) and a second part due to the confounding (second summand).

H.3 Quantifying confounding

Two indicators for confounding strength are proposed: i) a correlative, and ii) a structural indicator.

Definition 8 (correlative strength of confounding)**.** *The correlative strength of confounding gives the degree to which the confounder* U *contributes to the covariance between* $\mathbf{X}$ *and* Y*.*

$$\gamma := \frac{\|\Sigma_{\mathbf{X}U}\|^2}{\|\Sigma_{\mathbf{X}Y}\|^2 + \|\Sigma_{\mathbf{X}U}\|^2} \tag{70}$$

The following indicator for confounding strength, which measures the deviation of the estimable $\hat{\mathbf{a}}$ from the genuine causal parameter $\mathbf{a}$, is proposed

Definition 9 (structural strength of confounding)**.**

$$\kappa := \frac{\left\|c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}\right\|^2}{\|\mathbf{a}\|^2 + \left\|c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}\right\|^2}, \tag{71}$$

$$\kappa \in [0, 1]. \tag{72}$$

Note that from (69) and a normalizing condition

$$\mu_{A,v}(\mathbb{R}) = \|v\|^2$$

(eq. (10) in (Janzing and Schölkopf, 2018a)), one knows $\|\hat{\mathbf{a}}\|^2 \approx \|\mathbf{a}\|^2 + \left\|c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}\right\|^2$. Therefore, one can rewrite κ as

$$\kappa \approx \frac{\left\|c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}\right\|^2}{\|\hat{\mathbf{a}}\|^2} = \frac{\|\hat{\mathbf{a}} - \mathbf{a}\|^2}{\|\hat{\mathbf{a}}\|^2}. \tag{73}$$

In words, κ is the share of the influence of U on $\mathbf{X}$ of the overall strength of the association between Y and $\mathbf{X}$. Another interpretation: κ is the deviation of $\hat{\mathbf{a}}$ from $\mathbf{a}$ relative to the sum of squared length of $\hat{\mathbf{a}}$.

Note that the contribution of u to the covariance between $\mathbf{X}$ and Y is determined by the product $c\mathbf{b}$. As a consequence, rescaling c by some factor and $\mathbf{b}$ by its inverse leaves γ unaffected. Similarly, (a more sophisticated) rescaling of c and $\mathbf{b}$ leaves κ unaffected. The regimes with (i) large c and small $\mathbf{b}$ and with (ii) small c and large $\mathbf{b}$ can be thought of as two extremes on a continuum where knowing the value of U (i) hardly reduces the uncertainty about $\mathbf{X}$ or (ii) significantly reduces the uncertainty about $\mathbf{X}$. To capture these different regimes, JS propose an additional parameter that measures the explanatory power of U for $\mathbf{X}$,

$$\eta := tr(\Sigma_{\mathbf{XX}} - tr(\Sigma_{\mathbf{XX}|u})) = tr(\Sigma_{\mathbf{XX}}) - tr(\Sigma_{\mathbf{EE}}) = \|\mathbf{b}\|^2. \tag{74}$$

H.4 Estimating confounding

The vector-induced spectral measure of $\Sigma_{\mathbf{XX}}$ with respect to $\hat{\mathbf{a}}$ can be approximated by a normalized two parametric probability measure, $\nu_{\kappa,\eta}$, which decomposes into a causal part and a confounding part. The relative share of causal and confounding parts in that decomposition is given by κ. The algorithm proceeds by finding the normalized measure closest to (computable) $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$. The parameter constellation that minimizes the distance tells us the relative confounding strength.

How do JS do that? They show that $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$ asymptotically depends on four parameters (two of which, $\Sigma_{\mathbf{XX}}$ and $\hat{\mathbf{a}}$, can be estimated). Based on this insight, they formalize a two-parametric family of probability measures $\nu_{\kappa,\eta}$ such that it converges to $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$ up to a normalizing factor with high probability as the dimensionality of $\mathbf{X}$ increases:

$$\frac{1}{\|\hat{\mathbf{a}}\|^2}\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}} - \nu_{\kappa,\eta} \to 0 \text{ (weakly in probability)} \tag{75}$$

where

$$\nu_{\kappa,\eta} := (1-\kappa)\,\nu^{\text{causal}} + \kappa\,\nu_{\eta}^{\text{confounded}}. \tag{76}$$

We inspect each part in turn.

1. ν^{causal} is the hypothetical spectral measure that would be obtained in the absence of confounding. Following (67), it is defined as

$$\nu^{\text{causal}} := \mu_{\Sigma_{\mathbf{XX}}}^{\text{Tr}} \tag{77}$$

since, in the absence of confounding, the spectral measure induced by $\mathbf{a}$ should be equivalent to the tracial spectral measure of $\Sigma_{\mathbf{XX}}$ (up to a normalizing factor). The rotation-invariant prior on $\mathbf{a}$ is needed to justify this approximation. Point 1 in Postulate 1 is the formal argument. It relies on Lemma 4, which explicitly calls for a rotation-invariant prior.

2. To define the corresponding confounding part, JS propose an approximation to the spectral measure of $\Sigma_{\mathbf{XX}}$ induced by the vector $\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}$. Recall that $\mathbf{b}$ has generic orientation relative to $\Sigma_{\mathbf{EE}}$, see eq. (68). However, both $\mathbf{b}$ as well as $\Sigma_{\mathbf{EE}}$ are unknown. These two unknowns correspond to two steps that are important for constructing this approximation.

   (a) The eigen decomposition of $\Sigma_{\mathbf{EE}}$ reads QM_EQ^{-1} where $M_E := \mathrm{diag}(\lambda_1^E, \ldots, \lambda_d^E)$ with $\lambda_1^E > \cdots > \lambda_d^E$ eigenvalues of $\Sigma_{\mathbf{EE}}$. Although $\mathbf{b}$ is unknown, one does know that it is generic relative to $\Sigma_{\mathbf{EE}}$. Therefore, we can replace $\mathbf{b}$ with a vector that is 'particularly generic', namely $\mathbf{g} := (1, \ldots, 1)^\top/\sqrt{d}$, which satisfies

   $$\mu_{M_E,\mathbf{g}} = \mu_{M_E}^{\mathrm{Tr}}.$$

   Therefore, one can approximate the spectral measure of $\Sigma_{\mathbf{XX}}$ induced by the vector $\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}$ by spectral measure of $M_E + \eta\mathbf{g}\mathbf{g}^\top$ induced by $(M_E + \eta\mathbf{g}\mathbf{g}^\top)\sqrt{\eta}\mathbf{g}$. This construction is still not feasible as M_E, which contains the eigenvalues of $\Sigma_{\mathbf{EE}}$, is unobserved.

   (b) JS resort to a result stating that spectral measures are close in high dimensions:

   $$\mu_{\Sigma_{\mathbf{XX}}}^{\mathrm{Tr}} \approx \mu_{\Sigma_{\mathbf{EE}}}^{\mathrm{Tr}},$$

   see their Lemma 4. Therefore, one can approximate M_E with

   $$M_X = \mathrm{diag}(\lambda_1^X, \ldots, \lambda_d^X)$$

   and $\lambda_1^X > \cdots > \lambda_d^X$ eigenvalues of $\Sigma_{\mathbf{XX}}$.

   Putting these two steps together, JS define a rank-one perturbation of M_X as

   $$T := M_X + \eta\mathbf{g}\mathbf{g}^\top,$$

   compute the spectral measure of T induced by vector $T^{-1}\mathbf{g}$, and define

   $$\nu_\eta^{\mathrm{confounded}} := \frac{1}{\|T^{-1}\mathbf{g}\|^2}\mu_{T,T^{-1}\mathbf{g}}. \tag{78}$$

H.5 Algorithmic implementation

The algorithmic implementation proposed in JS (and used in Algorithm 1) is as follows. First, compute the observed vector-induced spectral measure $\mu_{\hat{\Sigma}_{\mathbf{XX}},\hat{\mathbf{a}}}$ where $\hat{\mathbf{a}}$ is the least-squares parameter vector resulting from a regression of Y on $\mathbf{X}$. Compute empirical counterparts of ν^{causal} in eq. (77) and $\nu_\eta^{\mathrm{confounded}}$ in eq. (78). Those empirical counterparts can then be used to compute eq. (76) for specific choices of κ and η, call the resulting quantity

$\hat{\nu}_{\kappa,\eta}$. The algorithm finds κ and η that minimize the distance between the observed vector-induced spectral measure and $\hat{\nu}_{\kappa,\eta}$. Since $\mu_{\hat{\Sigma}_{\mathbf{XX}},\hat{\mathbf{a}}}$ and $\hat{\nu}_{\kappa,\eta}$ have the same support (namely, the eigenvalues of $\hat{\Sigma}_{\mathbf{XX}}$), minimizing the distance between the respective weight vectors is sufficient. Call w the weight vector of $\mu_{\hat{\Sigma}_{\mathbf{XX}},\hat{\mathbf{a}}}$ and w' the weight vector of $\hat{\nu}_{\kappa,\eta}$.

JS propose computing the distance $D(w, w')$ by first smoothing w and w' using a Gaussian kernel and then taking the ℓ_1 norm of the resulting distance matrix,

$$D(w, w') := \left\| K(w - w') \right\|_1 , \tag{79}$$

with

$$K(\lambda_i, \lambda_j) := \exp\Big(- \frac{(\lambda_i - \lambda_j)^2}{2\sigma^2} \Big)$$

where $\sigma = 0.2$.

Finally, using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm, as implemented in the general purpose optimization `R` package `optim`, they search for κ and η that minimize $D(w, w')$. See also the description of the algorithm on page 12 of JS.

Appendix I. Janzing and Schölkopf (2018b) in a nutshell

The model analyzed in Janzing and Schölkopf (2018a) (see previous Appendix H) has a one-dimensional confounder. In a follow-up paper, Janzing and Schölkopf (2018b) show how to estimate a degree of confounding in models with a high-dimensional confounder. This Appendix reproduces the arguments in that paper and does not contain new results.

I.1 The set-up

The multi-dimensional confounder $\mathbf{Z}$ consists of $l \geq d$ independent sources each having unit variance and zero mean. It influences d-dimensional covariates $\mathbf{X}$ and one-dimensional Y as follows:

$$\mathbf{X} = M\mathbf{Z} \tag{80}$$

$$Y = \mathbf{X}^\top \mathbf{a} + \mathbf{c}^\top \mathbf{U} \tag{81}$$

where M is a $d \times l$-dimensional mixing matrix, $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{c} \in \mathbb{R}^l$, $\mathbf{a}$ contains the causal effect of $\mathbf{X}$ on Y.

This model induces the following covariance matrices

$$\Sigma_{\mathbf{XX}} = MM^\top \tag{82}$$

$$\Sigma_{\mathbf{X}Y} = MM^\top \mathbf{a} + M\mathbf{c} \tag{83}$$

and the parameter vector after projecting with least-squares in the population $\hat{\mathbf{a}}$ is given by

$$\hat{\mathbf{a}} = \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{X}Y} = \mathbf{a} + M^{-T}\mathbf{c} \tag{84}$$

where M^{-T} is the transpose of the pseudoinverse of M.

The degree of confounding is defined (excactly as in Janzing and Schölkopf, 2018a) as

Definition 10 (structural strength of confounding)**.**

$$\kappa := \frac{\|\hat{\mathbf{a}} - \mathbf{a}\|^2}{\|\mathbf{a}\|^2 + \|\hat{\mathbf{a}} - \mathbf{a}\|^2} \in [0, 1]. \tag{85}$$

The critical assumptions are the $\mathbf{a}$ and $\mathbf{c}$ are drawn from a rotation-invariant prior distribution (i.e. a distribution that is invariant with respect to orthogonal transformations). The following theoretical derivations rely on $\mathbf{a}$ and $\mathbf{c}$ drawn from a Gaussian distribution:

$$\mathbf{a}_i \sim \mathcal{N}(0, \sigma_a^2), \tag{86}$$

$$\mathbf{c}_i \sim \mathcal{N}(0, \sigma_c^2). \tag{87}$$

A key insight in JSb is that κ can then be approximated as a function of (estimable) $\Sigma_{\mathbf{XX}}$ and θ, the fraction of σ_a^2 and σ_c^2:

$$\theta := \frac{\sigma_c^2}{\sigma_a^2}. \tag{88}$$

Specifically, using some concentration of measure results for large d (details below), κ can be approximated as

$$\kappa \approx \frac{\frac{1}{d}\mathrm{tr}(\Sigma_{\mathbf{XX}}^{-1})\sigma_c^2}{\frac{1}{d}\mathrm{tr}(\Sigma_{\mathbf{XX}}^{-1})\sigma_c^2 + \sigma_a^2} = \frac{\frac{1}{d}\mathrm{tr}(\Sigma_{\mathbf{XX}}^{-1})\theta}{\frac{1}{d}\mathrm{tr}(\Sigma_{\mathbf{XX}}^{-1})\theta + 1} \tag{89}$$

Thus, JSb show how to infer θ, which can then be plugged in eq. (89) to calculate κ.

I.2 Estimating θ

Given the model described in the previous sub-section, the distribution of $\hat{\mathbf{a}}$ depends on unobserved M and l. The goal here is to construct a generating model for $\hat{\mathbf{a}}$ that only depends on observable quantities $\Sigma_{\mathbf{XX}}$ and d while generating the same distribution as that of $\hat{\mathbf{a}}$. It is shown in Theorem 1 in JSb that generating $\mathbf{b} \in \mathbb{R}^d$ by drawing each component from a standard Gaussian distribution and setting

$$\hat{\beta}_{\text{gen}} := \sqrt{\sigma_a^2 \mathbf{I} + \sigma_c^2 \Sigma_{\mathbf{XX}}^{-1}}\mathbf{b} \tag{90}$$

generates vectors with the same distribution as $\hat{\mathbf{a}}$ in the generating process of the previous subsection, i.e. vectors defined in eq. (84).

The generating model induces a distribution for $\frac{\hat{\mathbf{a}}}{\|\hat{\mathbf{a}}\|}$, which is the Haar measure on the orthogonal group (i.e. the uniform distribution on the unit sphere S^{d-1}), under the map

$$\mathbf{b} \mapsto \frac{\sqrt{\mathbf{I} + \theta\Sigma_{\mathbf{XX}}^{-1}}\mathbf{b}}{\left\|\sqrt{\mathbf{I} + \theta\Sigma_{\mathbf{XX}}^{-1}}\mathbf{b}\right\|}. \tag{91}$$

JSb show that the log probability density of the normalized parameter vectors,

$$v := \frac{\hat{\mathbf{a}}}{\|\hat{\mathbf{a}}\|} \tag{92}$$

is a function of θ and has the following form:

$$\log p_\theta(v) = \frac{1}{2}\left[\log\det(\mathbf{I} + \theta\Sigma^{-1}_{\mathbf{XX}}) - d\log\left\langle v, (\mathbf{I} + \theta\Sigma^{-1}_{\mathbf{XX}})^{-1}v\right\rangle\right] \tag{93}$$

If one had access to many samples from $\hat{\mathbf{a}}$, one could use them to maximize their log likelihood using (93), and thereby infer θ (which could then be used to calculate κ using eq. (89)). Usually, one does not have access to many samples from $\hat{\mathbf{a}}$. Remarkably, the authors proceed to show that having a large d is sufficient to estimate θ (and, therefore, κ via eq. (89)). Intuitively, drawing one high-dimensional vector $\hat{\mathbf{a}}$ is equivalent to drawing d components of that vector independently with respect to an appropriate basis.

Appendix J. Results when using Janzing and Schölkopf (2018b) to estimate κ

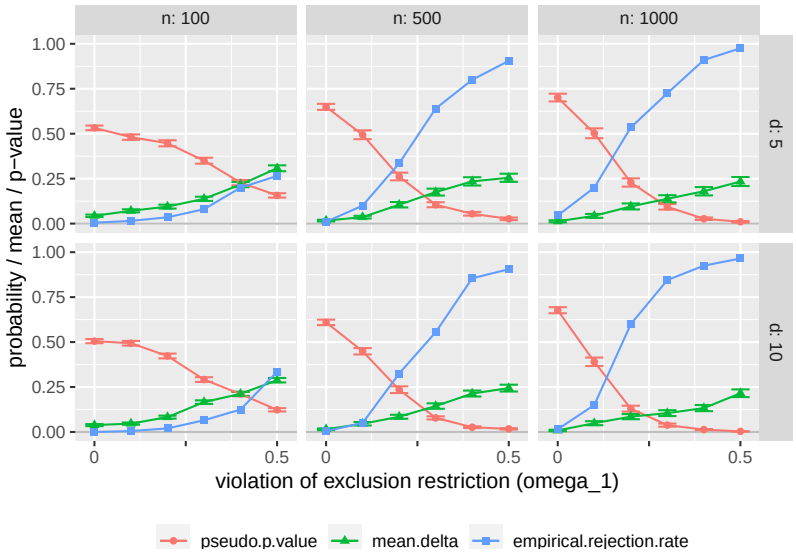

Figure 14: *Simulation results: pseudo-p-values, δ_B, and empirical rejection rate as a function of ω_1. Same figure as Figure 5, expect that JSb is used to estimate κ.*

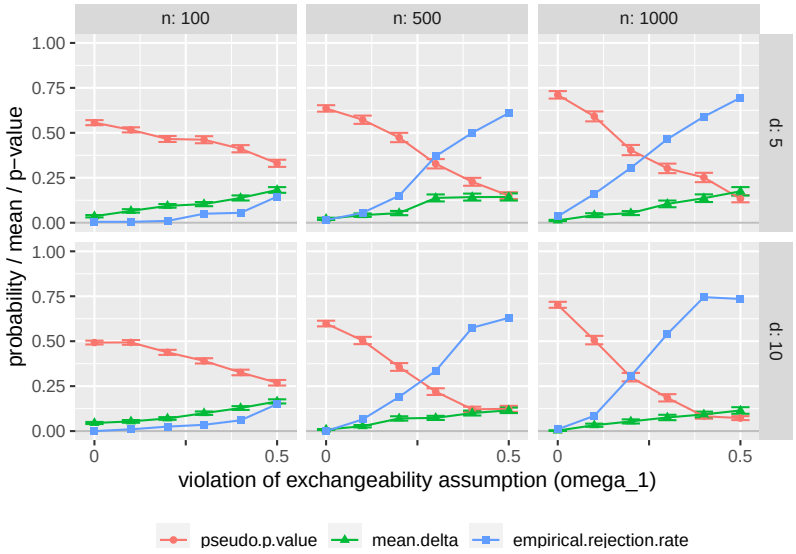

Figure 15: *Simulation results: pseudo-p-values, δ_B, and empirical rejection rate as a function of ω_1. Same figure as Figure 12, expect that JSb is used to estimate κ.*

	test results for different sets of covariates		
	$\{R, S\}$	$\{S\}$	$\{R\}$
p-value (JS)	0.18	0.29	0
p-value (JSb)	0.19	0.32	0
no. of covariates	29	6	23
no. of observations	3612	3612	3612

Table 3: *This Table shows results of the empirical application, based on Card (1995) for both methods to calculate κ. S denotes the set of covariates implicitly defining the subgroups in which the instrument is valid according to Card. R contains all remaining covariates (for details see main text). Consistent with Card's argument, the null hypothesis of instrument validity cannot be rejected when all covariates are included; see column $\{R, S\}$. Similarly, when only the six covariates S are included the instrument validity can also not be rejected; see column $\{S\}$. Dropping all variables S and keeping only those in R, the test rejects instrument validity.*

	test results for different methods to estimate κ	
	JS	JSb
p-value	0.52	0.58
no. of covariates	7	7
no. of observations	245	245

Table 4: *This Table shows results of the empirical application based on Acemoglu (2008). The null hypothesis of instrument validity cannot be rejected regardless of which method to estimate κ (JS or JSb) is used.*

Appendix K. Relation to knockoff procedure by Candes et al. (2018)

The knockoff procedure by Candes et al. (2018) is a method to do variable selection in high-dimensional settings while controlling the false discovery rate.[13] The central idea is to construct so-called "knockoff" variables $\widetilde{\mathbf{X}}$ based on the original variables $\mathbf{X}$ that are, conditional on those original variables, unrelated to the target variable Y,

$$\widetilde{\mathbf{X}} \perp\!\!\!\perp Y|\mathbf{X}. \tag{94}$$

The authors show that using variable selection methods (such as ridge regression) on a model that contains both $\widetilde{X}_i$ *and* X_i as explanatory variables, picks only those features X_i that are indeed related to Y while controlling a specified false discovery rate.

Both the knockoff procedure as well as Algorithm 2 are designed to construct variables (the knockoff and synthetic treatment variable, respectively) that resemble their original counterparts in specific ways. However, there is a slight difference between what the synthetic treatment variable and what the knockoff variables are supposed to fulfill. To illus-

13. We appreciate an anonymous reviewer's encouragement to compare our method with the work by Candes et al. (2018).

	test results for different sets of covariates		
	$\{R, S\}$	$\{S\}$	$\{R\}$
p-value	0.25	0.294	0.002
no. of covariates	29	6	23
no. of observations	3612	3612	3612

Table 5: *Results of the empirical application to Card (1995), equivalent of Table 1 but with the knockoff procedure used to generate the synthetic treatment variable T_s. S denotes the set of covariates implicitly defining the subgroups in which the instrument is valid according to Card. R contains all remaining covariates (for details see main text). Consistent with Card's argument, dropping all variables S and keeping only those in R, the test rejects instrument validity. Using the knockoff procedure instead of Algorithm 2 hardly affects results.*

trate that main difference, consider the relation of the constructed variable ($\widetilde{X}_i$ and T_s, respectively) with the output variable Y. In the knockoff procedure, the constructed variables are independent of Y conditional on the original variables, see expression (94). This is not what we want to achieve with the construction of the synthetic variable in Algorithm 2: To make the argument that $b_{T_s} = 0$ (where b_{T_s} is the coefficient of T_s in a regression of Y on $\{\mathbf{X}, T_s\}$, see Appendix A), we want $T_s \perp\!\!\!\perp Y|\mathbf{X}$, unconditionally on T_{IV}. Put differently, using the knockoff procedure constrains T_s to fulfill $T_s \perp\!\!\!\perp Y|T_{\text{IV}}$, which is not what we want.

The relation of the knockoff procedure to our approach of creating a synthetic variable is subtle and intriguing. Therefore, we describe how our results change if we use the knockoff procedure to create the synthetic treatment variable T_s instead of Algorithm 2. Note that a knockoff variable $\widetilde{X}_i$ and its original counterpart X_i fulfill $\text{Cov}(\widetilde{X}_i, X_j) = \text{Cov}(X_i, X_j) \forall j \neq i$, which is exactly the covariance structure that we want the synthetic treatment variable T_s and its original counterpart T_{IV} to fulfill, namely $\text{Cov}(T_s, X_j) = \text{Cov}(T_{\text{IV}}, X_j)$ where X_j are the remaining control covariates.

We reproduce the main simulation results in Table 5 and the results of both empirical applications when replacing line 6 of Algorithm 1 with the "fixed-X", equicorrelated knockoff construction to generate a knockoff version of T_{IV} (for details on the knockoff construction, please see Candes et al., 2018). Table 5 and Figure 16 show that the results discussed in the main text hardly change when using the knockoff procedure instead of Algorithm 2 to generate T_s.

Regarding the empirical application to Acemoglu et al. (2008), also when using the knockoff procedure to generate T_s, we cannot reject the null hypothesis of instrument validity with a p-value of 0.55.

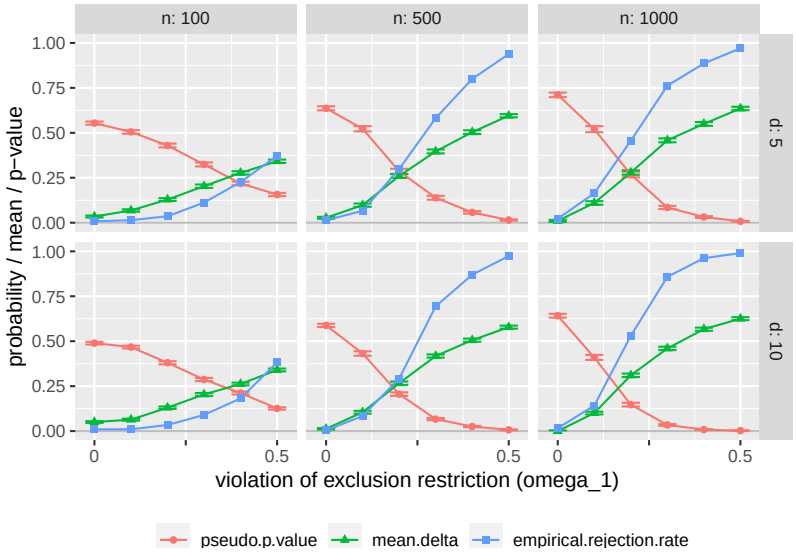

Figure 16: *Simulation results: pseudo-p-values, δ_B, and empirical rejection rate as a function of ω_1. Same figure as Figure 5, expect that knockoff procedure instead of Algorithm 2 is used to generate T_s.*

References

Daron Acemoglu, Simon Johnson, James A Robinson, and Pierre Yared. Income and democracy. *American Economic Review*, 98(3):808–42, 2008.

Daron Acemoglu, Suresh Naidu, Pascual Restrepo, and James A Robinson. Democracy does cause growth. *Journal of Political Economy*, 127(1):47–100, 2019.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. ISSN 1537274X. doi: 10.1080/01621459.1997.10474074.

Michel Besserve, Naji Shajarisales, Bernhard Schölkopf, and Dominik Janzing. Group invariance principles for causal generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 557–565. PMLR, 2018a.

Michel Besserve, Rémy Sun, and Bernhard Schoelkopf. Counterfactuals uncover the modular structure of deep generative models. *International Conference on Learning Representations*, 2020b.

Richard Blundell and Joel Horowitz. A Non Parametric Test of Exogeneity. *Review of Economic Studies*, 74(4):1035–1058, 2007. ISSN 0034-6527. doi: 10.1111/j.1467-937X.2007.00458.x. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-937X.2007.00458.x/full`.

Christoph Breunig. Goodness-of-fit tests based on series estimators in nonparametric instrumental regression. *Journal of Econometrics*, 184(2):328–346, 2015.

Christoph Breunig. Specification testing in nonparametric instrumental quantile regression, 2018.

Christoph Breunig and Xiaohong Chen. Adaptive, rate-optimal hypothesis testing in nonparametric iv models. *Cowles Foundation Discussion Papers*, (2671), 2020.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

David Card. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In Loizos Christofides, Kenneth Grant, and Robert Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, pages 201–222. University of Toronto Press, Toronto, 1995. URL `http://www.nber.org/papers/w4483`.

Elizabeth C Cooksey. Using the national longitudinal surveys of youth (nlsy) to conduct life course analyses. In *Handbook of life course health development*, pages 561–577. Springer, Cham, 2018.

Angus Deaton. Instruments, randomization, and learning about development. *Journal of economic literature*, 48(2):424–55, 2010.

Vanessa Didelez, Sha Meng, Nuala A Sheehan, et al. Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25(1):22–40, 2010.

Zheng Fang and Juwon Seo. A projection framework for testing shape restrictions that form convex cones. *Econometrica*, 89(5):2439–2458, 2021.

Ragnar Frisch, Trygve Haavelmo, T.C. Koopmans, and J. Tinbergen. Autonomy of Economic Relations. *League of Nations Memorandum*, 1938.

Patrick Gagliardini and Olivier Scaillet. A specification test for nonparametric instrumental variable regression. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, (128):151–202, 2017.

A. N. Gorban and I. Y. Tyukin. Blessing of dimensionality: Mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2118), 2018. ISSN 1364503X. doi: 10.1098/rsta.2017.0237.

Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems*, 34:28233–28248, 2021.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

James J Heckman and Edward Vytlacil. *Structural equations, treatment effects, and econometric policy evaluation*, volume 73. 2005. ISBN 4030008526. doi: 10.1111/j.1468-0262.2005.00594.x.

Michael Holmes and Mark Caiola. Invariance properties for the error function used for multilinear regression. *PLoS ONE*, pages 1–25, 2018.

Martin Huber and Giovanni Mellace. Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints. *Review of Economics and Statistics*, 97(2):638–647, 2015. ISSN 1725-2806. doi: 10.1162/REST.

Samuel P Huntington. Democracy's third wave. *Journal of Democracy*, 2(2):12–34, 1991.

Guido W Imbens and Joshua D Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994. doi: 10.1.1.363.2755.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. ISSN 00189448. doi: 10.1109/TIT.2010.2060095.

Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018a.

Dominik Janzing and Bernhard Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. *International Conference on Machine Learning*, pages 2245–2253, 2018b.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012. ISSN 00043702. doi: 10.1016/j.artint.2012.01.002.

Toru Kitagawa. A Test for Instrument Validity. *Econometrica*, 83(5):2043–2063, 2015. ISSN 0012-9682.

Jeremy Labrecque and Sonja A Swanson. Understanding the assumptions underlying instrumental variable analyses: a brief review of falsification strategies and related tools. *Current epidemiology reports*, 5(3):214–220, 2018.

Jan Lemeire and Dominik Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23(2):227–249, 2013. ISSN 09246495. doi: 10.1007/s11023-012-9283-1.

Shuo Li, Liuhua Peng, and Yundong Tu. Testing independence between exogenous variables and unobserved errors. *Econometric Reviews*, pages 1–32, 2022.

Seymour Martin Lipset. Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review*, 53(1):69–105, 1959.

N Gregory Mankiw, David Romer, and David N Weil. A contribution to the empirics of economic growth. *The Quarterly Journal of Economics*, 107(2):407–437, 1992.

Ismael Mourifié and Yuanyuan Wan. Testing Local Average Treatment Effect Assumptions. *Review of Economics and Statistics*, 99(2):638–647, 2017. ISSN 1725-2806. doi: 10.1162/REST.

Emily Oster. Unobservable Selection and Coefficient Stability : Theory and Evidence Unobservable Selection and Coefficient Stability : Theory and Evidence. *Journal of Business and Economic Statistics*, 37(2):187—-204, 2019. ISSN 0735-0015. doi: 10.1080/07350015.2016.1227711. URL `https://doi.org/10.1080/07350015.2016.1227711`.

Judea Pearl. *Causality: Models, Resoning, and Inference*. Cambridge University Press, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press (Open-access publication), Cambridge, Massachusetts, London, England, 2017. ISBN 9780262037310.

John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, pages 393–415, 1958.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2002.